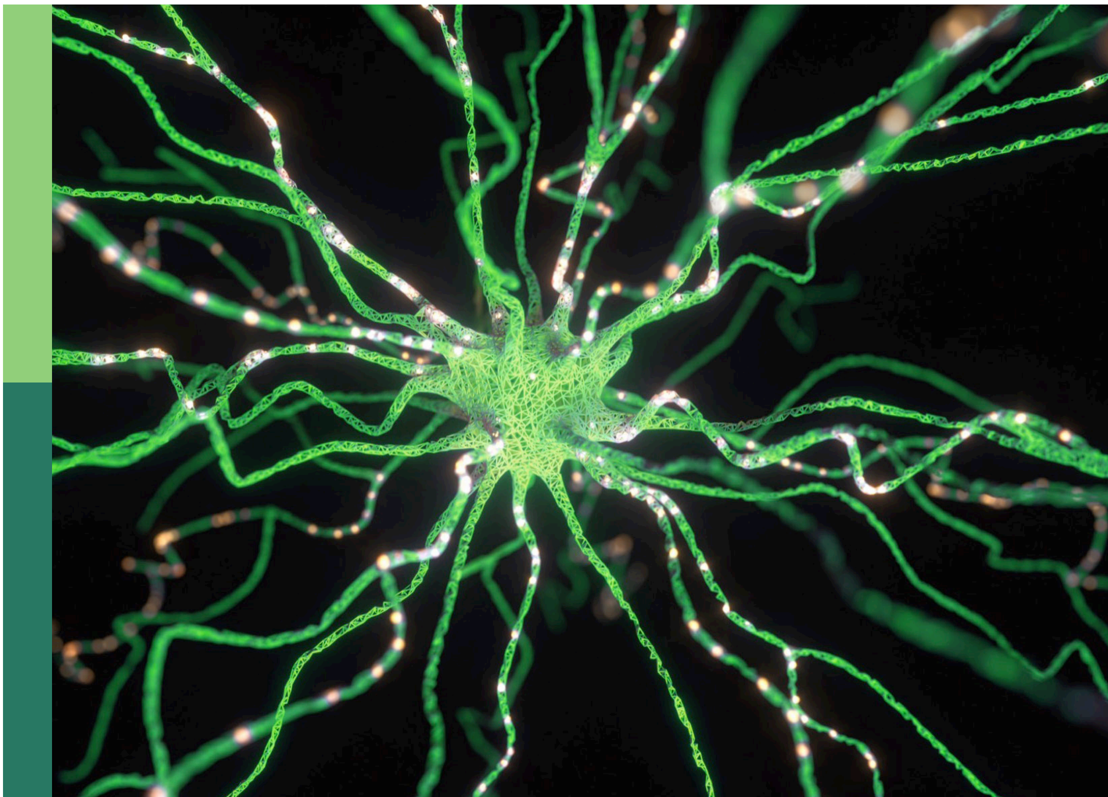
frontiers | Research Topics

Dynamic neural networks for robot systems: Data-driven and model-based applications

Edited by
Long Jin, Predrag S. Stanimirovic and
Sendren Sheng-Dong Xu

Published in
Frontiers in Neurorobotics



ISSN 1664-8714
ISBN 978-2-8325-5201-8
DOI 10.3389/978-2-8325-5201-8

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Dynamic neural networks for robot systems: Data-driven and model-based applications

Topic editors

Long Jin — Lanzhou University, China

Predrag S. Stanimirovic — University of Niš, Serbia

Sendren Sheng-Dong Xu — National Taiwan University of Science and Technology, Taiwan

Citation

Jin, L., Stanimirovic, P. S., Xu, S. S.-D., eds. (2024). *Dynamic neural networks for robot systems: Data-driven and model-based applications*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-5201-8

Table of contents

05 Advances on intelligent algorithms for scientific computing: an overview
Cheng Hua, Xinwei Cao, Bolin Liao and Shuai Li

26 An advanced bionic knee joint mechanism with neural network controller
Changxian Xu, Zhongbo Sun, Chen Wang, Xiujun Wu, Binglin Li and Liming Zhao

35 Location estimation based on feature mode matching with deep network models
Yu-Ting Bai, Wei Jia, Xue-Bo Jin, Ting-Li Su and Jian-Lei Kong

51 Active fault-tolerant anti-input saturation control of a cross-domain robot based on a human decision search algorithm and RBFNN
Ke Wang, Yong Liu and Chengwei Huang

70 An improved model for target detection and pose estimation of a teleoperation power manipulator
Li Xie, Jiale Huang, Yutian Li and Jianwen Guo

84 Adversarial robustness in deep neural networks based on variable attributes of the stochastic ensemble model
Ruoxi Qin, Linyuan Wang, Xuehui Du, Pengfei Xie, Xingyuan Chen and Bin Yan

98 Dual graph convolutional networks integrating affective knowledge and position information for aspect sentiment triplet extraction
Yanbo Li, Qing He and Damin Zhang

110 Occlusion facial expression recognition based on feature fusion residual attention network
Yuekun Chen, Shuaishi Liu, Dongxu Zhao and Wenkai Ji

121 A novel approach to attention mechanism using kernel functions: Kerformer
Yao Gan, Yanyun Fu, Deyong Wang and Yongming Li

134 Based on cross-scale fusion attention mechanism network for semantic segmentation for street scenes
Xin Ye, Lang Gao, Jichen Chen and Mingyue Lei

146 Robust control for a tracked mobile robot based on a finite-time convergence zeroing neural network
Yuxuan Cao, Boyun Liu and Jinyun Pu

159 Monocular catadioptric panoramic depth estimation via improved end-to-end neural network model
Fei Yan, Lan Liu, Xupeng Ding, Qiong Zhang and Yunqing Liu

170 Vision-force-fused curriculum learning for robotic contact-rich assembly tasks
Piaopiao Jin, Yinjie Lin, Yaoxian Song, Tiefeng Li and Wei Yang

182 Research on automatic pilot repetition generation method based on deep reinforcement learning
Weijun Pan, Peiyuan Jiang, Yukun Li, Zhuang Wang and Junxiang Huang

195 SafeCrowdNav: safety evaluation of robot crowd navigation in complex scenes
Jing Xu, Wanruo Zhang, Jialun Cai and Hong Liu

205 Cross-modal self-attention mechanism for controlling robot volleyball motion
Meifang Wang and Zhange Liang

221 Dense captioning and multidimensional evaluations for indoor robotic scenes
Hua Wang, Wenshuai Wang, Wenhao Li and Hong Liu

232 Voting based double-weighted deterministic extreme learning machine model and its application
Rongbo Lu, Liang Luo and Bolin Liao

241 Vehicle re-identification method based on multi-attribute dense linking network combined with distance control module
Xiaoming Sun, Yan Chen, Yan Duan, Yongliang Wang, Junkai Zhang, Bochao Su and Li Li

256 Context-aware SAR image ship detection and recognition network
Chao Li, Chenke Yue, Hanfu Li and Zhile Wang

271 A data-driven acceleration-level scheme for image-based visual servoing of manipulators with unknown structure
Liuyi Wen and Zhengtai Xie

285 Advancing autonomy through lifelong learning: a survey of autonomous intelligent systems
Dekang Zhu, Qianyi Bu, Zhongpan Zhu, Yujie Zhang and Zhipeng Wang

frontiers | Frontiers in Neurorobotics

TYPE Review
PUBLISHED 21 April 2023
DOI 10.3389/fnbot.2023.1190977

OPEN ACCESS

EDITED BY
Long Jin,
Lanzhou University, China

REVIEWED BY
Yinyan Zhang,
Jinan University, China
Vasilios N. Katsikis,
National and Kapodistrian University of Athens, Greece
Weibing Li,
Sun Yat-Sen University, China

*CORRESPONDENCE
Bolin Liao
✉ bolinliao@jsu.edu.cn
Shuai Li
✉ shuai.li@oulu.fi

†These authors share first authorship

RECEIVED 21 March 2023
ACCEPTED 06 April 2023
PUBLISHED 21 April 2023

CITATION
Hua C, Cao X, Liao B and Li S (2023) Advances on intelligent algorithms for scientific computing: an overview.
Front. Neurorobot. 17:1190977.
doi: 10.3389/fnbot.2023.1190977

Advances on intelligent algorithms for scientific computing: an overview

Cheng Hua[1†], Xinwei Cao[2†], Bolin Liao[1*] and Shuai Li[3,4*]

[1]College of Computer Science and Engineering, Jishou University, Jishou, China, [2]School of Business, Jiangnan University, Wuxi, China, [3]Faculty of Information Technology and Electrical Engineering, University of Oulu, Oulu, Finland, [4]VTT Technical Research Centre of Finland, Oulu, Finland

The field of computer science has undergone rapid expansion due to the increasing interest in improving system performance. This has resulted in the emergence of advanced techniques, such as neural networks, intelligent systems, optimization algorithms, and optimization strategies. These innovations have created novel opportunities and challenges in various domains. This paper presents a thorough examination of three intelligent methods: neural networks, intelligent systems, and optimization algorithms and strategies. It discusses the fundamental principles and techniques employed in these fields, as well as the recent advancements and future prospects. Additionally, this paper analyzes the advantages and limitations of these intelligent approaches. Ultimately, it serves as a comprehensive summary and overview of these critical and rapidly evolving fields, offering an informative guide for novices and researchers interested in these areas.

KEYWORDS

neural networks, intelligent systems, robotic, dynamic systems, optimization algorithms and strategies

1. Introduction

In recent years, the fields of computer science and communication electronics have undergone rapid growth and development, primarily due to the increasing interest in techniques that can enhance the performance of systems. The advancement of technologies such as neural networks, intelligent systems, optimization algorithms, and strategies has resulted in significant progress and created new opportunities and challenges in the areas of artificial intelligence, automation, and data science.

Neural networks, a potent machine learning algorithm, have garnered considerable attention due to their ability to solve intricate problems in diverse fields, such as speech recognition, image processing, and reinforcement learning. Inspired by the human brain's structure, neural networks consist of interconnected layers of nodes or "neurons" that process input data and generate output predictions. The primary advantage of neural networks stems from their self-learning capability, which enables them to assimilate knowledge from vast amounts of data and make accurate predictions without explicit programming. Consequently, they find extensive applications in domains where traditional programming is arduous and cumbersome. Additionally, neural networks can handle non-linear relationships between inputs and outputs, rendering them highly suitable for complex non-linear problems that are challenging to solve with linear models. However, neural networks also possess certain limitations, such as: (1) Black-box nature: Neural networks are often regarded as black-box models due to the challenge in comprehending how they arrive at their prediction outcomes. Consequently, diagnosing and rectifying errors in the model can be difficult; (2) Overfitting: Neural networks are susceptible to overfitting, which implies that they may perform well on the training data but poorly on new and unseen data. This can be mitigated by utilizing regularization techniques, but it continues to pose a challenge. (3) Training complexity: Neural networks are computationally intensive and time-consuming to train, particularly for large and complex datasets. In general, neural networks are potent tools in the realm of machine learning and have demonstrated considerable potential in solving intricate problems (Xiao et al., 2018b; Long et al., 2022; Peng and Liao, 2022; Liao et al., 2023). With sustained research efforts and continued development, they may offer even greater utility across a broad range of applications.

Intelligent systems have evolved into a pervasive and indispensable element of modern society. These systems utilize artificial intelligence and electronic communication technology to provide solutions for diverse applications, ranging from self-driving cars to home automation systems (Khan et al., 2022e). The widespread implementation of intelligent systems can be attributed to the steady advancement of technologies such as design, recognition, detection, prediction, and evaluation. Furthermore, the exceptional performance of intelligent system components, including communication systems and oscillators, assumes a crucial role. Communication systems are indispensable for transmitting data and commands between distinct components of the system (Zhang et al., 2022a), while oscillators provide accurate timing and synchronization to ensure the proper operation of the system (Jin et al., 2017a).

Optimization represents a fundamental challenge in multiple domains, entailing the identification of the optimal solution to a problem that complies with prescribed criteria and constraints. Optimization algorithms and strategies seek to automate this process and attain the optimal solution efficiently. Over time, diverse optimization algorithms have been developed, which can be broadly categorized into classical and metaheuristic approaches. Classical methods rely on mathematical techniques such as linear programming (Hu et al., 2019a), quadratic programming (Xiao, 2016; Xiao et al., 2019c), and dynamic programming (Lv et al., 2018; Liao et al., 2019), while metaheuristic methods are more heuristic and often inspired by natural phenomena (Sun et al., 2016; Khan et al., 2020a; Qu et al., 2020; Zhang et al., 2022b). Optimization methods and strategies play a critical role in the efficacy and competitiveness of various fields (Khan et al., 2021). For instance, optimization technologies can be employed to enhance the performance of machines or systems while reducing costs. Furthermore, optimization methods can have a favorable impact on society by improving the efficiency of public services and infrastructure, and addressing societal challenges such as poverty, inequality, and climate change. Overall, optimization methods and strategies constitute a crucial aspect from all perspectives.

This paper aims to present a comprehensive survey of three areas of research: neural networks, intelligent systems, and optimization algorithms and strategies. The basic principles, techniques, recent advances, and future directions of these intelligent methods will be explored in depth. This paper will provide a detailed examination of the models, algorithms, and applications used in each of these research fields. Furthermore, the advantages and limitations of these technologies will be thoroughly analyzed and discussed to aid readers in understanding and evaluating these intelligent methods. The structure of this paper is presented as follows. In Section 2, we categorize neural network models into real-valued and complex-valued types, and examine the activation function, robustness, and convergence of these models. Moreover, this section illustrates the relevant application domains of neural networks, including linear systems, non-linear systems, and robotic and motion planning. Section 3 discusses the pertinent technologies and components of intelligent systems, comprising system design, recognition, and detection methods, prediction and evaluation methods, and intelligent communication systems and oscillators. In Section 4, we explore bio-inspired optimization algorithms and optimization strategies and systems. Finally, Section 5 provides concluding remarks.

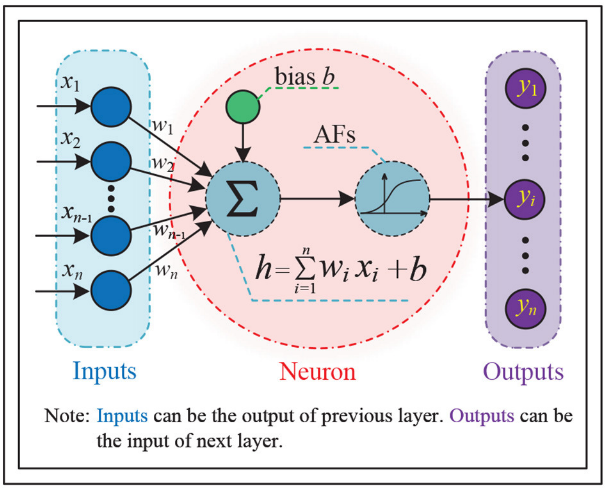

x_1
x_2
x_{n-1}
x_n
w_1
w_2
w_{n-1}
w_n
bias b
AFs
Σ
$h=\sum_{i=1}^{n} w_i x_i + b$
y_1
y_i
y_n
Inputs
Neuron
Outputs
Note: Inputs can be the output of previous layer. Outputs can be the input of next layer.

FIGURE 1
General structure of a single neuron in the most basic type of neural networks, where x_i denotes the ith input of the neuron, w_i is the corresponding weight, y_i represents the ith output of the neuron, and the activation functions (AFs) can be linear or non-linear.

2. Neural networks

2.1. Background

Neural networks are mathematical models that simulate the processing of complex information by the human brain's nervous system, based on the principles of neural networks in biology. These models abstract the structure of the brain and its response mechanism to external stimuli, and are represented by a large number of interconnected nodes (called neurons) with specific output functions (called activation functions or AFs). Connections between nodes represent weighted values (called weights) for signal transmission, allowing neural networks to simulate human memory. The network's output depends on its structure, connections, weights, and activation functions, which are typically approximations of algorithms, functions of nature, or logical strategies. Figure 1 illustrates the structure of a single neuron in the most basic type of neural network.

The neural network model has gained significant attention across various scientific domains due to its distinctive properties, which are as follows:

- **Self-learning and self-adaptive ability:** The neural network model is capable of adjusting its network structure parameters automatically when exposed to changes in the external environment (such as new training samples), to achieve the desired output corresponding to a specific input. Compared to traditional expert systems with fixed reasoning, neural network models are more adaptable and mimic the thinking style of the human brain.
- **Non-linearity:** Many real-world problems are viewed as non-linear complex systems, while neural networks store information in the number of neurons and connection weights, allowing for various non-linear mappings.
- **Fault-tolerance and robustness:** The distributed nature of information storage in neural network models ensures that local damage to the model moderately weakens the operation of the neural network without producing catastrophic errors. Moreover, neural networks can handle incomplete or noisy data, possess generalization function, and exhibit strong fault tolerance.
- **Computational parallelism and distributed storage:** The structural features of neural networks result in natural parallelism. Each neuron can perform independent operations and processing based on the received information and output the result. Different neurons in the same layer can perform operations simultaneously and then transmit to the next layer for processing. As a result, neural networks can take advantage of parallel computing to increase their operation speed significantly. Neural networks use distributed storage to represent information. By distributing the activation signals on the network neurons in response to the input information, the features are accurately remembered in the connection weights of the network through training and learning, enabling the neural network to make quick judgments when the same patterns are input again.

In the preceding subsection, we have acquired an initial comprehension of the fundamental architecture and characteristics of neural network models. In the following analysis, we will examine the models in greater detail from the standpoint of their various categories, problem-solving approaches, and practical applications.

2.2. Real-valued neural network model

Real-valued neural networks are a type of machine learning model that can process continuous data, making them highly versatile and effective in various domains, such as computer vision, natural language processing, and signal processing. For example, in image recognition, real-valued neural networks can take the pixel values of a digital image as input and produce the corresponding label as output. In stock price prediction, these networks can model historical stock data and provide trend predictions for future stock prices. In voice recognition, acoustic signals can be transformed into textual output through the use of real-valued neural networks. The activation function (AF) is a crucial component of the neural network architecture as it enables the transformation of the input into an output. Without an AF, the neural network can only represent linear functions. The addition of a non-linear AF allows the neural network model to achieve non-linear transformations from input to output, thereby enhancing its expressive power.

2.2.1. Neural network model with linear AF

Let us first consider the neural network model with a linear AF. In this case, the gradient, or derivative, of the neural network remains constant for each iteration, making it difficult for the model to capture complex information from the data. However, linear AF is still suitable for simple tasks that require high interpretability. In their study (Ding et al., 2014), the authors proposed a class of static recurrent neural network (SRNN) models with linear activation function and time-varying delays. To assess the stability of the SRNN model, they introduced a new Lyapunov-Krasovskii function and derived improved time delay-dependent stability conditions in the form of linear inequalities. They then provided numerical results that are consistent with the theoretical findings by specifying the SRNN model parameters. In another study (Zhang et al., 2019), the authors extended the original linearly activated fixed-parameter neural network to a linearly activated varying-parameter neural network model, where the parameter is chosen as $\zeta(t) = \alpha + \alpha^t$. Subsequently, Xiao et al. proposed an improved varying parameter neural network model (Xiao et al., 2020c). The parameter value of this model is

$$\zeta(t) = \begin{cases} \alpha + t^{\alpha}, & \text{if } 0 < \alpha \leq 1, \\ \alpha^2 + 2t\alpha + \alpha^{t+2}, & \text{if } \alpha > 1, \end{cases}$$

which can better meet the needs of the model hardware implementation.

The integration of various neural network approaches has garnered significant interest in addition to the investigation of individual neural network models. A novel strategy combining gradient-based neural networks (GNNs) and zeroing neural networks (ZNNs) was proposed in Dai et al. (2022) to solve dynamic matrix inversion online. The proposed strategy incorporates fuzzy adaptive control, which allows for adaptive adjustment by regulating the fuzzy factors based on real-time residual error values. The authors demonstrate the global convergence and efficacy of this GNN-ZNN model based on fuzzy control through theoretical analysis and numerical experiments. Different papers have employed various neural network models for the same problem, each with their own unique characteristics (Zhang et al., 2019; Xiao et al., 2020c; Dai et al., 2022). Therefore, exploring how to effectively combine the strengths of multiple neural network models in different scenarios is an important area of research. Fuzzy control theory, a mathematical theory dealing with fuzziness, is based on the concept of fuzzy sets and has been widely studied, including applications such as fuzzy inference (Zeng et al., 2022) and fuzzy Petri nets (Zhou et al., 2015, 2018a,b, 2019). These fuzzy control methods offer guidance for extending single neural networks to multi-neural networks.

2.2.2. Neural network model with non-linear AF

Non-linear AFs are a crucial element of neural networks, contributing to their expressive power and learning capability, leading to superior performance in handling complex tasks. Based on convergence properties, non-linear AFs can be categorized into two types: general AFs and finite-time convergent AFs.

(i) **General AFs:** In recent years, several studies have proposed neural network models with non-linear activation functions for solving a variety of problems. For example, in Jian et al. (2020), a class of neural network models was presented for solving the time-varying Sylvester equation, where the authors considered three different types of non-linear activation functions and provided a detailed theoretical derivation to validate the convergence performance of the proposed models. Similarly, Lei et al. proposed an integral structured neural network model with a coalescent activation function optimized for the solution of the time-varying Sylvester equation (Lei et al., 2022). For non-convex and non-linear optimization problems, an adaptive parameter convergence-differential neural network (CDNN) model with non-linear activation functions was proposed in Zhang et al. (2018d), and the authors verified the global convergence and robustness of the model by theoretical analysis and numerical experiments. Non-linear activation functions are also widely used in many fields, such as wheeled mobile robot control (Xiao et al., 2017b), surgical endoscopic robot control (Li et al., 2022b), and distributed collaborative networks (Zhang et al., 2018a).

(ii) **Finite-time convergent AFs:** Contrary to the general non-linear activation functions with infinite time convergence, the activation functions with finite time convergence facilitate fast convergence of neural network models, with a time upper bound. In Xiao et al. (2018a), the authors proposed a neural network model for online solution of Lyapunov equations in non-linear systems. The model's fast convergence was achieved by incorporating non-linear activation functions, and an upper bound on the model's time convergence was established via theoretical analysis as

$$\text{Time}_{\text{up}} < \frac{\alpha_1 + \beta_1}{\alpha_1\beta_1(1-\zeta)}\max\left\{|r^-(0)|^{(1-\zeta)}, |r^+(0)|^{(1-\zeta)}\right\},$$

where α_1 and β_1 are scale factors, $\zeta \in (0, 1)$, $r^+(0) = \max\{R(0)\}$, and $r^-(0) = \min\{R(0)\}$ with $R(0)$ denotes the initial value of the error function $R(t)$. Finally, the stability and finite-time properties of the model were confirmed in an application involving the control of a six-link robotic arm. In a similar vein, Xiao et al. developed an accelerated convergence recurrent neural network (RNN) model (Xiao, 2017a, 2019) for time-varying matrix square root finding (Zhang et al., 2015), and provided a time upper bound for the convergence of the model, which is expressed as

$$\text{Time}_{\text{up}} = \frac{\alpha_2}{\beta_2(\alpha_2-\gamma)}\ln\frac{\beta_2 A(0)^{(\alpha_2-\gamma)/\alpha_2}+\lambda}{\lambda},$$

where β_2 and λ are scale factors, $\alpha_2 > \gamma$ and all are odd integers, and $A(0)$ is a random initial value of the error matrix. For dynamic non-linear optimization problems (Liao et al., 2015; Xiao and Lu, 2019; Lu et al., 2020), the authors proposed a sign-bi-power AF and use it for dynamic neural network model design, and express the upper bound of the model convergence time mathematically as

$$\text{Time}_{\text{up}} < \max\left\{\frac{|\boldsymbol{k}_0^+|^{(1-\alpha_3)}}{\beta_3(1-\alpha_3)}, \frac{|\boldsymbol{k}_0^-|^{(1-\alpha_3)}}{\beta_3(1-\alpha_3)}\right\}$$

where α_3 is the scale factor, $1 < \beta_3 < 1$, $\boldsymbol{k}_0^+$ and $\boldsymbol{k}_0^-$ represent the maximum and minimum initial values of the error vector $\boldsymbol{k}$, respectively. In order to account for the effects of rounding errors and external noise disturbances in practical problem solutions, Xiao and colleagues proposed a neural network model in Xiao et al. (2019d) with the capability to suppress noise and achieve predefined time convergence. The authors provided detailed theoretical proof of the robustness and finite-time convergence of the model. They also verified through numerical experiments that the model can still achieve finite-time convergence in the presence of external noise. In Liao et al. (2022a), a predefined time-convergent neural network model with harmonic-like noise suppression was designed for adaptively solving time-varying problems by leveraging the properties of harmonic signals. The burgeoning demand for real-time performance has become a critical requirement for many scientific, industrial, and commercial applications, such as computational biology, weather forecasting, autonomous vehicles, and financial analytics. This requirement is largely driven by the rapid progress in computer technology, including advances in hardware and software, which have enabled the processing of vast quantities of data in real-time (Tan and Dai, 2017; Dai et al., 2018; Tan, 2021; Li et al., 2022a). Real-time performance is essential for many time-sensitive applications, where delays or inaccuracies in processing can have severe consequences, such as in real-time monitoring of critical physiological signals or detecting anomalies in sensor data. Furthermore, real-time performance enables immediate feedback and adaptive decision-making, leading to increased efficiency and performance. In Zhang et al. (2022c), the authors proposed a unified GNN model for handling both static matrix inversion and time-varying matrix inversion with finite-time convergence and a simpler structure. As the authors conclude, compared with the existing GNN model and ZNN model dedicated to time-varying matrix inversion, the proposed unified GNN model has advantages in convergence speed and robustness to noise. At the same time, the authors further extend this GNN model for finding the dynamic Moore-Penrose inverses in real-time (Zhang et al., 2022d), and the paper concludes that this method does not require the time derivatives of the relevant dynamic matrices and has finite time convergence. In short, high-precision and low-complexity real-time solutions are a highly active area of research, with numerous open problems and opportunities for innovation in both fundamental algorithms and system-level optimizations.

To facilitate the reader's understanding, we present a list of the linear and non-linear activation functions discussed in Section 2 and provide a detailed description of each function in Table 1.

(1) Linear activation function (LAF):

$$\mathcal{A}(x) = x. \tag{1}$$

(2) Power activation function (PAF):

$$\mathcal{A}(x) = x^{\mu} \text{ with } \mu > 3 \text{ indicating an odd integer.} \tag{2}$$

TABLE 1 Details of various linear and non-linear activation functions.

AFs	Type	References
LAF (1)	Linear	(Ding et al., 2014; Zhang et al., 2019; Jian et al., 2020; Xiao et al., 2020c; Dai et al., 2022)
PAF (2)	Non-linear	(Jian et al., 2020)
BPAF (3)	Non-linear	(Zhang et al., 2018a; Lei et al., 2022)
PSAF (4)	Non-linear	(Zhang et al., 2018d)
HSAF (5)	Non-linear	(Xiao et al., 2017b; Li et al., 2022b)
SBPAF (6)	Non-linear & Finite-time convergence	(Xiao, 2017a, 2019; Xiao et al., 2018a, 2019d)
TSBPAF (7)	Non-linear & Finite-time convergence	(Liao et al., 2022a)

(3) Bipolar sigmoid activation function (BPAF):

$$\mathcal{A}(x) = (1 - \exp(-\mu x))/(1 + \exp(-\mu x)) \text{ with } \mu > 1. \tag{3}$$

(4) Power-sigmoid activation function (PSAF):

$$\mathcal{A}(x) = \begin{cases} x^{\mu}, & \text{if } |x| \geq 1, \\ \dfrac{1 - \exp(-\mu x)}{1 + \exp(-\mu x)} \cdot \dfrac{1 + \exp(-\mu)}{1 - \exp(-\mu)}, & \text{otherwise.} \end{cases} \tag{4}$$

(5) Hyperbolic sine activation function (HSAF):

$$\mathcal{A}(x) = (\exp(\mu x) - \exp(-\mu x))/2 \text{ with } \mu > 1. \tag{5}$$

(6) Sign-bi-power activation function (SBPAF) :

$$\mathcal{A}(x) = (|x|^{\mu} + |x|^{1/\mu})\text{sgn}(x)/2 \text{ with } 0 < \mu < 1, \tag{6}$$

thereinto,

$$\text{sgn}(x) = \begin{cases} 1, & \text{if } x > 0, \\ 0, & \text{if } x = 0, \\ \text{-}1, & \text{if } x < 0. \end{cases}$$

7) Tunable sign-bi-power activation function (TSBPAF):

$$\mathcal{A}(x) = \frac{1}{2}\rho_1|x|^{\mu}\text{sgn}(x) + \frac{1}{2}\rho_2 x + \frac{1}{2}\rho_3|x|^{1/\mu}\text{sgn}(x), \tag{7}$$

where $\mu \in (0, 1)$, ρ_1, ρ_2, and ρ_3 are greater than 1.

2.3. Complex-valued neural network model

In recent years, neural network-based machine learning techniques have found broad application in practical settings. Notably, the majority of current neural network models are designed for real-valued inputs, outputs, and weights. However, this raises the question of the existence and purpose of complex-valued neural network models. What are complex-valued neural network models, and why are they necessary? Complex-valued neural network models utilize complex numbers as inputs, outputs, and weights and are inspired by the natural properties of complex numbers and the existence of complex-valued neurons in biology. They are employed in specific application scenarios where the input and output data can be represented in complex form, and therefore, complex-valued neural networks can better describe and process these data. Compared to real-valued neural networks, complex-valued neural networks offer several advantages:

- They can better represent complex-valued data in the real world, such as sound waves and electromagnetic waves.
- They can achieve better results with a smaller network size due to the effectiveness of complex-valued weights in expressing correlations and symmetries in the data.
- They can better handle asymmetrical data by allowing for expression rotation and scaling, which can map asymmetric data into a more symmetric space.
- They can better handle phase information, which is important for complex-valued data, as traditional real-valued neural network models struggle to handle the phase information effectively.

Complex-valued neural networks have been extensively employed in image recognition, speech recognition, and natural language processing, and are currently under thorough investigation. In the following sections, we will delve into the complex-valued neural network model and scrutinize it through the lenses of noise-tolerance and finite-time convergence.

2.3.1. Noise-tolerance

The precision and robustness of neural network models can be adversely affected by computational rounding errors and external noise perturbations. Therefore, it is crucial for these models to possess the dual capability of solving problems and suppressing noise simultaneously.

In Xiao and Lu (2017), a complex-valued gradient neural network model was proposed for solving complex-valued linear matrix equations. This model has a simpler theoretical analysis and lower computational complexity compared to the widely used real-valued gradient-based neural network model. In Lei et al. (2020), the authors proposed a neural network model for computing the inverse of complex-valued time-varying matrices. The model's convergence in solving time-varying problems and its robustness against external noise disturbances were analyzed and validated. The effect of design parameters on the speed of model solving was also elucidated based on experimental results. Moreover, a complex-valued noise-resistant neural network model based on an integral-type design formulation was presented in Xiao et al. (2019f) for the same problem. The convergence and robustness of the model were verified through detailed analysis and proofs. The experiments considered various noise types, including constant noise, linear noise, bounded linear noise, harmonic noise, and exponential-type noise. The model proposed in this work has a better noise suppression effect compared to the traditional gradient-based neural network model. To further improve the noise tolerance of the neural network model, a complex-valued noise-tolerant neural network model with a double-integral structure was proposed in Liao et al. (2022b), which was capable of simultaneously solving the problem and suppressing the noise. The authors verified the robustness of the model under constant noise, linear polynomial noise, and quadratic polynomial noise via numerous theoretical analyses. According to the numerical experimental results, this model can achieve the effective suppression of constant noise, linear polynomial noise, and quadratic polynomial noise. In Ding et al. (2019b), Ding et al. proposed an improved complex-valued recurrent neural network (ICVRNN) model for solving the complex-valued time-varying Sylvester equation. This work gives a large number of theoretical proofs and experimental cases to analyze the effectiveness, convergence, and stability of the ICVRNN model. Additionally, the authors further extend this ICVRNN model to the solution of complex-valued linear equations (CVLEs) (Ding et al., 2018). As the authors conclude, the ICVRNN model has better performance for solving CVLEs compared to traditional neural network models. In addition, noise-tolerant complex-valued neural network models are widely used for solving many problems, such as matrix pseudo-inverse solving (Lei et al., 2019), robotics (Liao et al., 2022d), and non-linear optimization (Xiao et al., 2019a), etc.

2.3.2. Finite-time convergence

Finite-time convergence is a crucial characteristic of neural network models as it allows for achieving the desired level of performance in a shorter amount of time. Specifically, if a neural network model can attain convergence within a finite time, the parameter selection and tuning process can be expedited to obtain the desired results more quickly. The non-linear activation function used in complex-valued neural network models plays a pivotal role in achieving finite-time convergence. This function is based on the non-linear activation function in the real domain but generalized to the complex domain. Unlike its counterpart in the real domain, the complex-valued non-linear activation function operates on complex inputs and outputs, which enables better handling of the non-linear characteristics of complex-valued data.

In Li and Li (2013), Li et al. proposed two ways to generalize the AF from the real domain to the complex domain, as follows.

i) **Complex-valued AF Type I:**

$$\mathcal{F}(a + ib) = \mathcal{A}(a) + i\mathcal{A}(b),$$

where $\mathcal{F}(\cdot)$ is a complex-valued AF defined in an element-wise manner, and a and b denote the real and imaginary parts of the complex number $a + bi$, respectively.

ii) **Complex-valued AF Type II:**

$$\mathcal{F}(a + ib) = \mathcal{A}(\Upsilon) \diamond \exp(i\Theta),$$

where the symbol $\diamond$ denotes the multiplication of the corresponding subelements of two vectors or matrices (i.e., $\boldsymbol{c} \diamond \boldsymbol{d} = [c_j d_j]$ for real vectors $\boldsymbol{c} = [c_j]$ and $\boldsymbol{d} = [d_j]$), and $\Upsilon \in \mathbb{R}$ and $\Theta \in (-\pi, \pi]$ represent the modulus and argument of the complex number $a + bi$, respectively.

In Xiao et al. (2020b), the authors proposed two non-linear equivalent models for solving complex-valued problems. One model focused on the real and imaginary parts of the complex numbers, while the other was from the perspective of the modulus of the complex numbers. The authors introduced a non-linear activation function to ensure fast convergence and applied these models to solve the complex-valued Sylvester equation. Both models performed well, as reported by the authors. In Xiao et al. (2022b), the authors designed an arctan-type variable-parameter complex-valued neural network model with finite-time convergence. This model takes into account the reality that the convergence factor is time-varying in the actual hardware environment. During the solution process, the model can adjust its convergence scale parameters (CSPs). When the model achieves convergence, the CSPs converge to a constant greater than zero. The CSPs and finite-time upper bounds of this model are supported by theoretical analysis, as the authors conclude. The excellent performance of this model has been demonstrated in numerical experiments. Furthermore, the authors extended this variable-parameter neural network model to solve time-varying complex-valued matrix equations (Ding et al., 2018; Xiao et al., 2021b).

In Zhou et al. (2022), the authors aimed to improve the robustness and solution speed of complex-valued noise-resistant neural network models for practical problem-solving, while meeting the dual requirements of noise tolerance and real-time performance. To this end, the authors introduced non-linear activation to the model. In this work, the authors employed this improved model to solve the problem of trajectory tracking for manipulators, and the results demonstrate that this model can effectively suppress noise while meeting real-time requirements of the task. In another work (Xiao et al., 2021a), the authors utilized a complex representation to convert the quaternion-valued matrix into the corresponding time-varying complex-valued matrix (TVCVM), and then proposed a complex-valued neural network model to solve this TVCVM. The authors introduced a versatile non-linear-sign activation function to achieve the predefined time convergence of the model. According to the authors' summarized results, theoretical analysis provided an upper bound for the convergence time of this model. Finally, the authors applied this model to a mobile manipulator and demonstrated its good performance.

2.4. Neural networks for linear system solving

A linear system is characterized by the linear property, which states that the system response is homogeneous and additive, such that the output signal changes in proportion to the input signal of the system. Solving linear systems with neural networks is of significance as it enables fast processing via learning and optimization, particularly for problems that are difficult or computationally complex to solve by traditional methods. Compared to traditional solution methods, using neural networks to solve linear systems has the following advantages.

- **Strong solving ability:** It can handle large-scale, high-dimensional linear systems, where traditional methods may be computationally overloaded or numerically unstable.

- **Good adaptability:** It can adaptively learn the mapping relationship between input and output, this allows neural networks for more complex linear system solving.
- **High accuracy in solving:** It can improve the accuracy of the model by increasing the number of layers and neurons of the neural network, this makes the neural network applicable to the solution of linear systems with high accuracy requirements.

2.4.1. Linear equation

In many real-time applications, including control and signal processing, precise analysis and control of linear systems are crucial. To this end, various neural network models have been proposed for the online solution of time-varying linear systems. For instance, in Lu et al. (2019), the authors introduced a novel recurrent neural network (RNN) model for solving time-varying underdetermined linear systems while satisfying the constraints of state variables and residual errors. This work presented extensive theoretical analyses and numerical cases to demonstrate the effectiveness and validity of the proposed RNN model, which was further applied to control the PUMA560 robot under physical constraints. In Xiao et al. (2019b), the authors developed a neural network model for time-varying linear matrix equations and provided a theoretical analysis of the upper bound on the time convergence of the model. The study concluded that this model demonstrated exceptional performance in solving time-varying linear equations. Additionally, in Zhang et al. (2018b), the authors proposed a varying-gain RNN model for solving the linear system $H(t)J(t)K(t) = L(t)$, with the design parameters of the model being characterized by time-varying properties. The finite-time convergence of this model was also verified by theoretical analysis. In Xiao et al. (2019e), two non-linear neural network models were investigated for solving the dynamic Lyapunov equation $H^{\mathrm{T}}(t)J(t) + J(t)H(t) = -K(t)$, and the study noted that the solution outcomes of these models were independent of the choice of initial values. Similarly, in Xiang et al. (2018a), the authors proposed a discrete Z-type neural network (DZTNN) model for the same dynamic Lyapunov equation, which exhibited inherent noise tolerance and exact solution attainment under various types of noise. Additionally, various neural network models (Xiao, 2017b; Jin et al., 2019; Xiao and He, 2021; Lei et al., 2022; Han et al., 2023) have been put forward for solving the time-varying Sylvester equations $H(t)J(t) - J(t)H(t) = -K(t)$.

2.4.2. System of linear equations

The system of linear equations is a fundamental mathematical concept used in various fields as a powerful tool to solve practical problems due to its linearity, simultaneousness, infinite solutions, and suitability for multiple methods. In Xiao et al. (2022a), the authors proposed a neural network model with adjustable parameters and demonstrated its fast convergence speed, low upper limit of convergence time, and short parameter adjustment time. The study also applied the model to achieve synchronous control of chaotic systems and validated its effectiveness. The authors concluded that this model performed excellently. In Xiao et al. (2017a), a gradient-based dynamic model was proposed for the simultaneous solution of systems of linear equations. The authors demonstrated that the model had a zero error bound at convergence and provided an upper bound on the convergence time. Additionally, this class of dynamic models was extended to the online solution of complex-valued systems of linear equations (Xiao, 2015; Xiao et al., 2021b). To meet the requirements of high real-time and strong robustness in solving linear systems of equations in engineering practice, in Xiao et al. (2020a), the authors developed a dynamic control model with noise robustness for online solution of systems of linear equations. The paper designed a non-linear activation function with noise tolerance and added it to the dynamic control model. The authors theoretically analyzed the noise immunity, convergence, and robustness of the model. Furthermore, the authors applied the dynamic control model to the motion tracking of the robot, and the results demonstrated good performance in the elliptical path tracking control of the robot. In Katsikis et al. (2023), the authors proposed a dynamic neural network model, based on neutrosophic numbers and a neutrosophic logic engine, which exhibits superior performance compared to the traditional ZNN design. The primary objective of this model is to estimate the matrix pseudo-inverse and minimum-norm least-squares solutions of time-varying linear systems. The observed enhancement in efficiency and accuracy of the proposed model over existing techniques is attributed to the advantages of neutrosophic logic over fuzzy and intuitionistic fuzzy logic. The authors utilized neutrosphication, de-fuzzification, and de-neutrosophication instead of the conventional fuzzification and de-fuzzification methods. The efficacy of the proposed model was assessed through simulation examples and engineering applications in the domains of localization problems and electrical networks.

2.5. Neural networks for non-linear system solving

Non-linear systems present a significant challenge for modeling, analysis, and control because their output cannot be described simply by a linear relationship with the input, and their dynamics may exhibit complex behaviors such as chaos or periodicity. The study of non-linear systems is critical to many fields, including control engineering (Xiao et al., 2017b; Zhou et al., 2022), signal processing (Jin, 2014; Luo and Xie, 2017), dynamics analysis (Tan and Dai, 2016; Tan et al., 2017, 2019a; Lu et al., 2020), and communication systems (Jin and Yu, 2012; Jin and Fu, 2013; Jin et al., 2015b; Zhao et al., 2020; Xiang et al., 2022), owing to the following properties.

- **Abundant kinetic behavior:** Unlike linear systems, the kinetic behavior of non-linear systems can be very abundant and diverse. For example, they can generate chaotic phenomena, periodic oscillations, and stable immobile points, etc.
- **Better modeling of complex phenomena in the real world:** Many natural and social phenomena are non-linear, such as ecosystems, economies, and neural systems. Non-linear systems can simulate these phenomena and provide relevant behavioral information.

- **Available for control and optimization:** Non-linear control theory is an important tool for applying non-linear systems to control and optimize problems. For example, in robotics and industrial control, non-linear control enables highly accurate and efficient solving of tasks.

In particular, non-linear systems can exhibit sensitivity to initial conditions, bifurcations, and singularities, making them a rich area of investigation for researchers. Furthermore, non-linear systems are capable of representing a wide range of phenomena, including self-organization, emergence, and adaptation, which are not captured by linear models. Thus, developing effective methods for modeling, analysis, and control of non-linear systems remains an important area of research in many disciplines.

Neural network methods are a powerful tool for real-time parallel processing that can be utilized to solve challenging non-linear systems, particularly for situations in which an analytical solution is elusive. These methods have found application in various domains, including non-linear control problems (Xiao et al., 2019g; Li et al., 2020c; Jia et al., 2021), non-linear differential equations (Zhang et al., 2017, 2018d; Liao et al., 2021), and non-linear optimization problems (Liu et al., 2016; Lan et al., 2017; Xiao et al., 2019a; Zhang et al., 2020).

2.5.1. System of non-linear equations

Non-linear systems frequently appear in real-world applications, and the online solution of systems of non-linear equations has been a subject of extensive research. One popular approach for solving such systems is through the use of neural network methods, which can be particularly useful when the analytical solution is difficult to obtain. In Xiao et al. (2019g), the authors proposed a class of recurrent neural network (RNN) models with finite-time convergence for solving systems of non-linear equations. The effectiveness of this RNN model was demonstrated through numerical simulations, and the model was extended to solve more complex non-linear systems, such as the motion tracking control of robotic manipulators. The authors concluded that this RNN model is highly feasible and applicable. Additionally, the authors constructed a discrete noise-resistant recurrent neural network (DNTRNN) model (Li et al., 2020c) based on the five-step finite difference method for the solution of non-linear systems of equations, and demonstrated the effectiveness of the DNTRNN model. In Liu et al. (2016), the authors proposed an RNN model for time-varying non-linear optimization, providing both continuous and discrete forms of the model. The paper concludes that both types of RNN models have superior noise immunity and convergence performance. In Zhang et al. (2018d), the authors designed and proposed a differential neural network with varying parameters and non-linear activation for solving non-convex optimization and non-linear problems online. The global convergence of this neural network model was proven through theoretical analysis, and the authors concluded that this neural network model performs well for solving non-convex and non-linear optimization problems in various numerical experiments.

2.5.2. Quadratic programming (QP)

The quadratic programming method is widely used in practice and is a powerful tool for solving practical problems, which has the following merits.

- **Can describe complex problems:** QP can describe numerous complex optimization problems, such as optimization problems with non-convex functions.
- **Available for constraint handling:** QP can handle optimization problems with constraints, such as inequality constraints, equation constraints, etc. This allows for a broader application of quadratic planning.
- **Extensive solving methods:** The solution methods of QP have been relatively mature, such as the gradient descent method, conjugate gradient method, and neural network method. These methods can be used in practice and can handle large-scale problems.
- **Global optimality:** QP guarantees global optimality for convex quadratic problems, which means that the solution found is guaranteed to be the best possible solution.

Neural network methods offer certain advantages in solving QP problems and are capable of solving large-scale QP problems. Additionally, they avoid the need for mathematical modeling and solving of problems in traditional algorithms. In Liao et al. (2021), the authors introduced neuro-dynamic methods for QP solving and pointed out the limitations of traditional neuro-dynamic methods in the presence of noise. Consequently, they proposed a predetermined time convergence neuro-dynamic method with inherent noise suppression and concluded that this method can achieve a fast and accurate solution to time-varying QP problems in noisy environments. In Zhang et al. (2020), the authors studied a power-type RNN (PT-RNN) model with varying parameters for time-varying QP and quadratic minimization (QM) solving under external perturbations. In this work, the authors provided a detailed design process of this PT-RNN model and analyzed the robustness and convergence of the model theoretically. Lastly, the authors used this model for venture investment and robot tracking. As the authors concluded, this PT-RNN model has great robustness and wide applicability. In Jia et al. (2021), the authors proposed a neural network approach based on an adaptive fuzzy control strategy for time-dependent QP solving. As summarized in the paper, this neural network method can automatically adjust the convergence parameters according to the residual error, which has better results compared with the traditional fixed-parameter neural network method. Similar to QP, non-linear programming (NLP) has also received much attention and is a powerful way to describe complex problems. In Katsikis and Mourtas (2021), the authors aimed to minimize portfolio insurance (PI) costs and presented a multi-period minimum-cost PI (MPMCPI) problem, which incorporates transaction costs, as a more practical version of the classical minimum-cost PI problem. The MPMCPI problem was formulated as a NLP problem, and the authors proposed an approach using intelligent algorithms to solve it. The efficacy of the proposed approach was evaluated using real-world data and compared with other meta-heuristic and commercial methods. The study results contribute to the optimization of portfolio insurance

TABLE 2 Comparison of the properties of neural network models in solving various types of problems.

Problems		Properties of NNs	References
Linear system	Linear equation	Finite-time convergence	(Xiang et al., 2018a; Zhang et al., 2018b; Lu et al., 2019; Xiao et al., 2019b,e; Xiao and He, 2021)
		Noise suppression	(Xiao, 2017b; Xiang et al., 2018a; Jin et al., 2019; Xiao et al., 2019b,e)
	System of linear equations	Finite-time convergent	(Xiao, 2015; Xiao et al., 2017a, 2022a)
		Noise suppression	(Xiao et al., 2020a)
Non-linear system	System of non-linear equations	Finite-time convergent	(Zhang et al., 2018d; Xiao et al., 2019g; Li et al., 2020c)
		Noise suppression	(Liu et al., 2016; Li et al., 2020c)
	Quadratic programming	Finite-time convergent	(Jia et al., 2021; Liao et al., 2021)
		Noise suppression	(Zhang et al., 2020; Liao et al., 2021)

NNs in this table indicate neural networks.

costs using intelligent algorithms and provide insights into the comparative performance of different approaches. Table 2 provides a summary of the works on neural network models for solving linear and non-linear systems.

2.6. Related applications

Neural networks are widely applied in various fields owing to their parallel computing capability, adaptive learning, and non-linearity. In this subsection, we provide a concise overview of the research on neural networks for redundant robot manipulators. A redundant robot manipulator is a robotic arm that has more degrees of freedom than required. The additional degrees of freedom are known as redundant degrees of freedom. Due to these redundant degrees of freedom, the robotic arm can be more flexibly adapted to different tasks and environments, as well as avoid obstacles or enhance motion performance by adjusting its posture. As a potent tool for real-time parallel processing, neural network models can be used for precise and flexible control of redundant robot manipulators (Xiao and Zhang, 2014; Zhang et al., 2014, 2018c; Liao and Liu, 2015; Jin et al., 2017b; Guo et al., 2018; Tan et al., 2019b; Xiao et al., 2019g; Li et al., 2020d, 2022b; Tang et al., 2022; Zhou et al., 2022). More specifically, neural networks can be used in two ways.

- **Inverse kinematic solving:** The redundant robot manipulator has additional degrees of freedom, and it can move the target position in multiple ways, thus the inverse kinematics needs to be solved to determine the best solution for the motion. Traditional inverse kinematics methods are susceptible to locally optimal solutions, while neural networks can obtain more accurate inverse kinematics solutions by autonomously adjusting the network structure and parameters.
- **Motion planning:** Redundant robot manipulators can use multiple postures to perform the same task, so the optimal sequence of postures needs to be determined for the optimal motion path. Adopting a neural network to solve the optimal posture sequence of the robot manipulator can achieve higher movement efficiency (Khan et al., 2022b).

2.6.1. Inverse kinematic solving

In Xiao and Zhang (2014), a dynamic neural network model is proposed for solving the inverse kinematics of mobile robot manipulators. The authors provided a theoretical analysis demonstrating the global convergence of the model to the inverse kinematic solution of the mobile robot manipulator, which is also supported by numerical experiments. The paper concludes that this dynamic model outperforms traditional gradient-based neural network models for the inverse kinematic solution of mobile robot manipulators. Liao et al. propose a bi-criteria pseudo-inverse minimization strategy for the redundancy problem of robot manipulators at the joint acceleration level (Liao and Liu, 2015), which can avoid high joint speeds of the manipulator. This method has been validated on a 4-degree-of-freedom robot manipulator and is found to perform well in solving the redundancy problem of robotic manipulators. Tang et al. used an enhanced planning scheme for redundant robot manipulator control (Tang et al., 2022), and a tuning strategy based on this scheme is found to achieve good results in the limit case. Zhang et al. propose a differential scheme with varying parameters for the joint-angle drift (J-AD) problem of redundant robot manipulators (Zhang et al., 2018c). The J-AD problem is formulated as a standard QP problem to be solved, and the authors validate this scheme through computer simulations and physical experiments, concluding that it performs well for solving the J-AD problem of redundant robot manipulators. Figure 2 depicts the schematic structure of a three-degree-of-freedom robot manipulator. In Zhang (2022), the authors discussed the problem of redundancy of manipulators in intelligent systems and designed a dynamic neural network with triple projections, called a tri-projection neural network (TPNN), which is developed for quadratic programs with a constraint on the state evolution of the neuron states. This paper concludes that the TPNN has advantages in fully employing the acceleration capability of the manipulator.

2.6.2. Motion planning

In Guo et al. (2018), a bi-criteria minimization scheme was proposed for motion planning of redundant robot manipulators, which incorporates joint velocity, joint acceleration, and joint angular constraints into the scheme. The authors design this

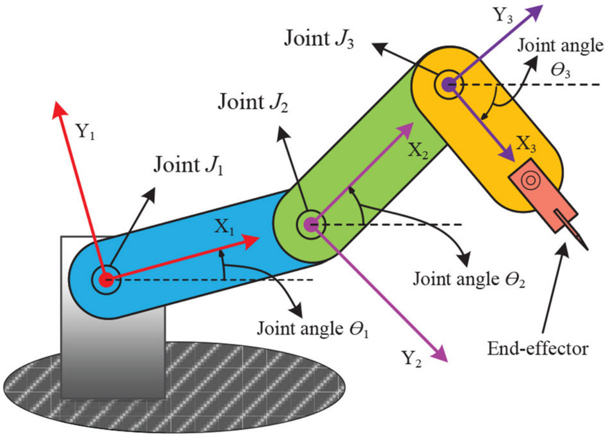


FIGURE 2
Schematic structure of a three-degree-of-freedom planar robot manipulator.

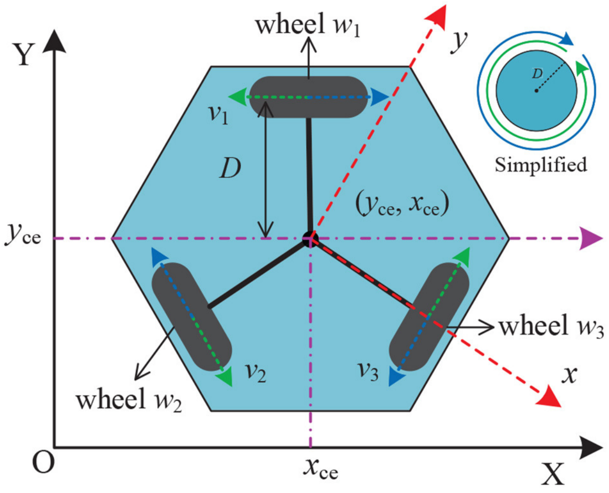


FIGURE 3
Geometric and kinematic model of an omnidirectional mobile wheeled robot, where (x_{ce}, y_{ce}) denotes the geometric center of the wheeled robot.

scheme based on the infinity norm acceleration minimization and minimum weighted velocity criterion. The authors evaluated the scheme through experimental simulations and physical validation, concluding that it is both excellent and physically realizable for redundant robot motion planning. In Jin et al. (2017b), the authors solved the distributed cooperative motion of redundant robot manipulators by reformulating it as a QP problem and designing a neural network model with noise tolerance for this QP problem. The authors validate this neural network model for the problem of the distributed cooperative motion of redundant robotic manipulators in noise-free and noise-containing environments, demonstrating its effectiveness on the PUMA560 redundant robot. Similarly, Li et al. investigated a neural network scheme with noise suppression and use it for redundant robot repetitive motion planning (Li et al., 2020d). The authors verified the effectiveness of this scheme on a four-link and a PA10 robot manipulator, concluding that its performance was superior to conventional motion planning schemes. In Zhang et al. (2014), a QP-based feedback control and motion planning scheme was designed and used for feedback control and motion planning of a mobile robot manipulator. The effectiveness of this scheme has been verified by dynamics analysis, and the authors conclude that it is reliable and superior for feedback control and motion planning of mobile robot manipulators. Figure 3 provides the geometric and kinematic model of an omnidirectional mobile wheeled robot.

2.7. Development directions and challenges

In recent years, neural networks have become a dominant technology in machine learning and artificial intelligence. They have achieved state-of-the-art results in various fields, such as image recognition, natural language processing, and game playing. However, neural networks still face several challenges, such as overfitting, data efficiency, and hardware constraints:. In this section, we will discuss the current state and future development directions of neural networks, as well as the challenges that may be faced in the future.

2.7.1. Development directions

Neural networks are expected to evolve in several directions in the future. There are some of the most promising directions:

- **Explainability:** One of the main challenges of neural networks is their lack of interpretability. It is often difficult to understand why a neural network makes a particular decision. Explainable AI (EAI) aims to address this issue by providing human-understandable explanations of the decisions made by neural networks. EAI is expected to become an essential aspect of AI in the future, especially in fields such as healthcare, finance, and autonomous systems.
- **Federated learning:** Federated learning is a distributed machine learning technique that allows multiple parties to collaboratively train a model without sharing their data. It is expected to become increasingly popular in the future due to its privacy-preserving nature. Federated learning can be used in various scenarios, such as personalized recommendation, fraud detection, and predictive maintenance.
- **Quantum neural networks:** Quantum neural networks (QNNs) are a type of neural network that utilizes quantum computation to process information. QNNs have the potential to outperform classical neural networks in various tasks, such as optimization, simulation, and cryptography. QNNs are expected to become increasingly important as quantum computing technology advances.

2.7.2. Challenges

Despite the many advancements in neural networks, they still face several challenges that need to be addressed in the future. There are some of the main challenges:

- **Overfitting:** Overfitting occurs when a neural network learns the noise in the training data instead of the underlying pattern. This can lead to poor generalization performance on new data.
- **Data efficiency:** Neural networks typically require a large amount of labeled data to achieve good performance. This can be a major bottleneck in real-world applications, especially in domains where data is scarce or expensive to obtain. One potential solution to this challenge is the development of transfer learning techniques that allow pre-trained models to be fine-tuned on smaller datasets.
- **Hardware constraints:** Neural networks require large amounts of computation and memory resources, which can be challenging to deploy on resource-constrained devices such as mobile phones and IoT devices. One potential solution is the development of hardware optimized for neural network computations, such as specialized processors and accelerators.

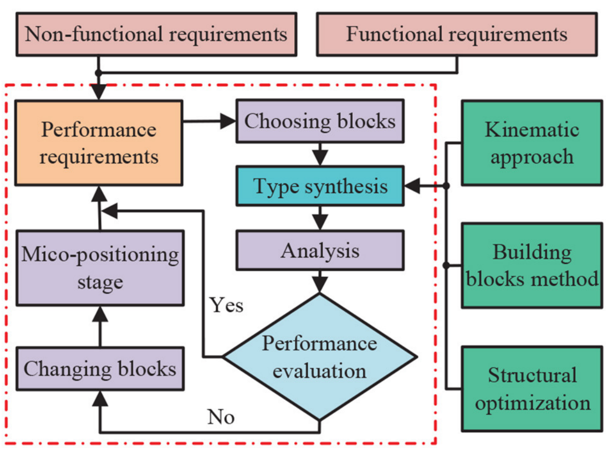


FIGURE 4
Detailed design framework of micro-positioning stages (MPSs), where the content in the red dotted box is the basic framework of MPSs.

3. Intelligent systems

An intelligent system is an automated system that leverages computer and artificial intelligence technology to enable intelligent decision-making, control, and management. It facilitates automatic control and optimization of various complex systems by collecting sensor data, processing information, and executing operations. Intelligent systems typically include the following components.

- **Sensors and actuators:** Used for sensing and controlling the state and operation of physical systems.
- **Data collection and processing module:** Used to collect, process and store sensor data, extract features of the system, and make decisions based on those features.
- **Decision and control algorithms:** Using artificial intelligence technology to analyze and process the data and achieve intelligent control of the system by control algorithms.

Intelligent systems have numerous applications, including industrial automation, intelligent medical care, intelligent home, and intelligent transportation. The wide range of potential applications suggests that the use of intelligent systems will become more widespread in the future, driving innovation and progress in numerous industries.

3.1. Design and control of intelligent systems

The design process plays a crucial role in determining the performance, reliability, maintainability, and scalability of intelligent systems. In this section, we will provide an overview of the current research on intelligent system design and control.

In Ding et al. (2021), the authors proposed an intelligent system combining a pseudo-rigid body approach and a constant force output mechanism for workpiece contact force control. In this work, the intelligent system was constructed as a mathematical model and provided a theoretical analysis to verify it. To obtain the optimal parameters and structure, a particle swarm optimization (PSO) method was used and experimentally verified by the authors. As the paper concludes, this intelligent system is excellent and generalizable. In Lan et al. (2016), the authors studied an observer design method for fractional-order one-sided Lipschitz intelligent systems. Also, the asymptotic stability of the full-order observer error system has been ensured by using an indirect Lyapunov method and an equivalent model. In Ding et al. (2019a), the authors investigated a design scheme for a reconfigurable planar micro-positioning stages (MPSs) based on different functional modules, and details the flexibility and functionality of this scheme were presented in the paper. Finally, the authors point out that the system provides a new idea for the design of MPSs. Facing the practical need for higher precision MPSs (Liao et al., 2022e), the authors proposed a novel assembly concept (both planar and spatial configurations) that further improves the flexibility and functionality of intelligent systems. Figure 4 presents the detailed design framework of MPSs.

In Ding et al. (2022), an intelligent system of constant force mechanism based on the combination of negative and positive stiffness was presented. In this work, the authors have modeled and validated the system. The results of this paper indicate that in numerical experiments, this intelligent system can achieve the required constant force output and was consistent with the theoretical results. In addition, a class of semi-interactive intelligent systems has been proposed for the creation of robotic dance works (Peng et al., 2015, 2016). The authors point out that this system was capable of self-adaptive and self-learning capabilities and has been validated on the NAO robot with good performance. Besides the above instances, the intelligent system also has widespread application scenarios, such as equipment processing control (Tang et al., 2015; Wu et al., 2021), substation management (Hu et al., 2021), and UAV collaborative control (Xu et al., 2022).

3.2. Identification and detection in intelligent systems

Recognition and detection technology, integrated with computer vision technology and machine learning algorithms, has become a critical component of intelligent systems. The fundamental concept of this technology is to analyze, process, and comprehend input images or videos to identify and detect target objects or events. By accomplishing automatic recognition, classification, localization, and tracking functions, recognition and detection technology can augment the intelligence and automation of intelligent systems. It has extensive applications, including but not limited to, facial recognition, autonomous driving, and security monitoring. The development of recognition and detection technology relies on advancements in computer vision, machine learning, and signal processing techniques, which are enabling the creation of more efficient and accurate recognition and detection algorithms. Ongoing research is focused on enhancing the robustness, accuracy, and real-time performance of recognition and detection technology, thereby expanding its applicability to a diverse range of real-world scenarios (Qin et al., 2017; Hu et al., 2019b; Zhuo and Cao, 2021; Niu et al., 2022).

3.2.1. Identification methods

In Zhuo and Cao (2022), the authors presented a novel approach for identifying damage in bolt connections of steel truss structures using sound signals. The proposed method employed support vector machine (SVM) classification, optimized with a genetic algorithm, to accurately recognize signals associated with bolt connection damage. The study demonstrated the effectiveness of SVM classification for signal recognition in structural health monitoring, specifically for detecting damage in bolt connections. In Wu et al. (2022c), a new scheme based on a low-strain pile integrity test and convolutional neural network (CNN) was proposed to identify concrete pile foundation defects with a remarkable accuracy of 94.4%. The authors described this method as more accurate, more reliable, and less destructive than traditional methods. Similarly, in Wu et al. (2022a), the authors proposed a method for the defect identification of foundation piles under layered soil conditions. In Tang et al. (2020), a human action recognition scheme was proposed, introducing and using the RGB-D image feature approach, which is a current research hotspot for effectively resisting the influence of external factors and improving the generalization ability of the classifier. The proposed scheme achieved excellent identification results on the public CAD60 and G3D datasets, utilizing three different patterns for human action feature extraction: The RGB modal information, based the histogram of oriented gradient (RGB-HOG), the depth modal information, based on the space-time interest points (D-STIP), and the skeleton modal information based on the joints' relative position feature (S-JRPF). In Xiang et al. (2018b), the authors identified Markov chains on trees (MCoT) through derivative constraints on the univariate distribution of sojourn time and/or hitting time, concluding that all MCoT can be identified using this method.

3.2.2. Detection methods

In Luo et al. (2020), the authors investigated a novel chaotic system and its associated signal detection method, demonstrating high detection accuracy and noise immunity in experimental studies. The effectiveness and feasibility of the proposed method were verified through theoretical analysis, circuit simulation, and FPGA implementation, highlighting its potential as a reliable solution for signal detection in chaotic systems. In Wu et al. (2022b), a deep learning-based system was proposed for structural damage detection of engineering steel beams, where the vibration signals were used to extract features and detected by CNN. The experimental results show that the accuracy of this detection method achieved 95.14%. The authors concluded that this method has superior performance for structural damage detection of engineering steel beams compared to the SVM method. Furthermore, in Chen et al. (2022a), the authors provided a comprehensive review of the techniques for detecting code duplication in software development, analyzing the advantages and disadvantages of each approach.

3.3. Prediction and evaluation in intelligent systems

Prediction and evaluation are crucial elements in intelligent systems, facilitating accurate decision-making, pattern identification, model optimization, and goal attainment. These components interact with other aspects of intelligent systems, including learning algorithms and models, prediction and planning, evaluation and optimization, and self-adaptation and self-optimization, leading to enhanced system optimization and development.

3.3.1. Prediction methods

Prediction is a crucial aspect of intelligent systems that can enable more informed decision-making, facilitate the discovery of regularities and patterns in data, optimize models, and support the attainment of system goals. Prediction can be achieved through the analysis of historical data to identify patterns and trends using intelligent systems. For instance, in Huang et al. (2022), the authors proposed a non-linear intelligent system for predicting the anti-slide pile top displacement (APTD) and identified multiple factors that affect the APTD. The proposed system was validated using four prediction methods, namely ELMAN, long short-term memory neural network (LSTM), support-vector regression (SVR), and maximal information coefficient-SVR (MIC-SVR), with results indicating superior performance in practical applications. Additionally, an integrated model based on wavelet transformation was introduced in Ding et al. (2013) for the prediction of both steady-state and dynamic-state network traffic. Low-frequency components were predicted using an improved gray theory, while the high-frequency components were predicted using a BP neural network algorithm, leading to increased prediction accuracy and reduced uncertainty. Moreover, an intelligent algorithm was introduced in Deng et al. (2019) for predicting the effective wind speed in wind turbines by considering the rotor speed, aerodynamic characteristics, and extreme learning machine. The authors reported that this algorithm is more efficient and accurate compared to traditional Kalman filter-based methods. Finally, an efficient search algorithm and optimization method were proposed in Song et al. (2020) to predict wind speed and extract the maximum wind energy.

3.3.2. Evaluation methods

Evaluation is a fundamental aspect of intelligent systems that allows for the assessment of the accuracy and performance of data, models, or decisions. During the evaluation process, the system compares actual values with ideal values to determine the accuracy and reliability of the model or decision. In the field of robotics, various methods have been proposed for the aesthetic evaluation of robotic dance movements. For instance, in Peng et al. (2022), the authors presented a method for aesthetic evaluation of robotic dance movements that employs key pose descriptors and integrated classifiers to train machine learning models. This method has been tested in a virtual environment and shown good performance. In Peng et al. (2019a), a brain-like intelligent system resembling the visual cognitive system of humans was proposed for the aesthetic evaluation of robotic dance poses. The system extracted features such as color, shape, and orientation and applied machine learning methods for evaluation. A computational framework for instantiating an intelligent evaluation method for robotic dance poses was presented in Figure 5. Similarly, in Li et al. (2020b), an automated method was proposed to evaluate the aesthetic level of robot dance movements by integrating multi-modal information. Features were extracted from visual and non-visual channels, and ten machine-learning algorithms were employed for evaluation, with the highest accuracy reaching 81.6%. Additionally, in Peng et al. (2019b), a feature fusion method was proposed for the automatic evaluation of robotic dance poses, which extracted four types of features, including color block, contour feature, region feature, and kinematic feature.

3.4. Intelligent communication systems

The intelligent communication system refers to a communication system that utilizes modern communication technology and artificial intelligence algorithms to dynamically adjust its parameters and structure based on varying communication needs, thereby achieving optimal communication performance and resource utilization efficiency. In this paper, we briefly describe three key aspects of the intelligent communication system: high-speed communication transmission methods, up-conversion mixer design, and spectrum sensing methods.

3.4.1. High-speed communication transmission

In Sun et al. (2022), the authors proposed a method to enhance the rate range and reduce power consumption in high-speed serial links by utilizing an adaptive continuous time linear equalizer (CTLE) and a half-rate decision feedback equalizer (DFE) with a hybrid filter and a current-integrating summer. The system was tested using 10 Gb/s PRBS7 signals transmitted through an 18-inch FR4 backplane, and the post-simulation results demonstrated a rate range of 6.25-10 Gb/s with excellent performance. In Zhang and Yang (2020), the authors proposed an adaptive CTLE based on slope detection and a half-rate inferred DFE with intermediate frequency compensation and a small amount of equalization for the middle frequency range. The measurements showed an effective equalization loss of 24 dB at Nyquist frequency with a clear eye diagram at 36 Gb/s. Both works provide solutions to the challenges of high-speed transmission and offer valuable insights into the design of receiver equalizers for high-speed serial links.

3.4.2. Up-conversion mixer design

In Chen et al. (2013), a folded up-conversion mixer was proposed by the authors, which employs a current reuse technique and achieves a conversion gain of 9.5 dB at a 1 V supply voltage while consuming only 258 μW of power. In Jin et al. (2014b), the authors presented a sub-harmonic up-conversion mixer that halves the required local oscillator frequency and achieves a higher conversion gain of 14.4 dB, albeit at the cost of increased power consumption of 1.65 mW at 1 V supply voltage. In Jin and Yu (2013), a current-reuse current-mirror-switch mixer was investigated by the authors, which features 8.5 dB conversion gain, 1.16 mW power consumption, lower supply voltage, higher linearity, and smaller chip area. All three works proposed novel mixers for wireless applications using 0.18-micron radio-frequency CMOS technology, with a focus on high performance, low power consumption, and small chip area, albeit with differences in specific technologies and performance metrics.

3.4.3. Spectrum sensing

In Yang et al. (2017), the authors investigated a multi-band spectral sensing method based on eigenvalue ratios, which employs random matrix theory to determine the distribution of new statistics solely in the presence of noise. This approach allows for the reliable establishment of theoretical thresholds and exhibits superior performance in small sample scenarios. In Lei et al. (2016), the authors introduced a blind broadband spectrum sensing algorithm based on principal component analysis. This algorithm transforms the wide-band spectrum sensing problem into a sequential binary hypothesis test utilizing a generalized likelihood ratio test, enabling simultaneous operation on all sub-bands and overcoming noise uncertainty issues. Both studies propose innovative approaches to addressing the multi-band spectral perception challenge, without requiring prior knowledge. The authors emphasized the practical significance of these methods for applications such as radio spectrum allocation, spectrum sharing, and dynamic spectrum access.

3.5. Intelligent oscillator systems

Intelligent oscillation systems are complex devices designed to generate controlled vibration signals that exhibit adjustable amplitude and frequency. Generally, these systems comprise several essential components, including a vibration source (e.g.,

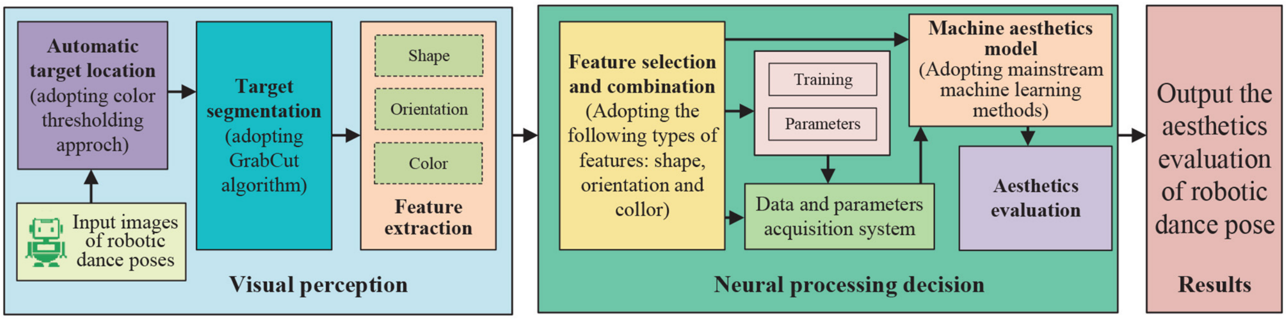

Automatic target location (adopting color thresholding approch)
Input images of robotic dance poses
Target segmentation (adopting GrabCut algorithm)
Shape
Orientation
Color
Feature extraction
Visual perception
Feature selection and combination (Adopting the following types of features: shape, orientation and collor)
Training
Parameters
Data and parameters acquisition system
Machine aesthetics model (Adopting mainstream machine learning methods)
Aesthetics evaluation
Neural processing decision
Output the aesthetics evaluation of robotic dance pose
Results

FIGURE 5
Computational framework for instantiating an aesthetic intelligence evaluation method for robotic dance poses. This evaluation method includes several components such as target localization, feature extraction, feature selection and combination, neural training, and decision-making.

a motor or piezoelectric device), a controller, sensors, and feedback loops. With a diverse range of applications, these systems have demonstrated their effectiveness in areas such as structural vibration control, acoustic and mechanical system testing, and medical devices.

3.5.1. Quadrature oscillator design

The quadrature oscillator is a passive oscillator that produces a sinusoidal wave with frequency and impedance determined by the inductor and capacitor values. This oscillator generates two orthogonal signals, sine and cosine waves, making it widely used in wireless communication systems. In Jin et al. (2015a), two variable frequency third-order quadrature oscillators (TOQOs) were proposed based on current differential transconductance amplifiers (CDTA). These TOQOs were completely resistorless and provided four quadrature current outputs at high output impedance terminals. In Jin and Liang (2013), a new resistorless current-mode quadrature oscillator based on CDTA was introduced, which provided two well-defined quadrature outputs at high-impedance terminals for easy cascading. Both works utilized CDTA for building the quadrature oscillator with the resistorless circuit, enabling monolithic integration, explicit orthogonal current outputs, direct cascading with other current-mode circuits, and controllable oscillation frequencies.

3.5.2. Quadrature voltage-controlled oscillator design

The quadrature voltage-controlled oscillator (QVCO) is an active oscillator that generates a sinusoidal wave, where the oscillation frequency is determined by an external control voltage. QVCO typically consists of two orthogonal oscillation circuits, which can vary the oscillation frequency by altering the phase difference between the two circuits. In Jin and Tan (2019), the authors proposed a novel low-voltage and low-power QVCO that is coupled by four P&N transistors, yielding a wide tuning range and low phase noise while consuming a meager 2.31 mW. In Jin (2018a), the authors introduced a novel QVCO architecture that employs four capacitors to achieve enhanced phase noise and reduced power dissipation compared to conventional designs. Furthermore, Jin et al. (2014a) developed a programmable current-mode multi-phase voltage-controlled oscillator (MPVCO) using cascaded first-order all-pass filters, which provides multiple outputs. These studies have introduced significant advancements in the design of voltage-controlled oscillators, resulting in enhanced performance, compact size, and reduced power consumption. These advancements are crucial for numerous applications in wireless communication systems.

3.5.3. Chaotic oscillator design

The chaotic oscillator is a non-linear dynamical system that exhibits complex, unpredictable behavior. It can be realized either through mathematical equations or physical circuits. In Jin (2018b), the authors proposed a novel digitally programmable multi-directional chaos oscillator (DPMDCO), which employs MOS switches for controlling the chaotic oscillation in three different directions. The DPMDCO achieves a compact size and low power consumption, making it suitable for practical applications. In Ouyang et al. (2022), a fully integrated chaotic oscillator (FICO) based on operational amplifiers and multipliers was presented. This system integrates all necessary circuit elements into a single chip, providing ease of implementation and compactness. Both DPMDCO and FICO were evaluated using the Cadence IC design tool, with DPMDCO consuming 99.5 mW at ± 2.5 V supply voltage and occupying 0.177 mm^2 of chip area, while FICO consumed 148 mW and had a larger chip area of 6.15 mm^2. These works demonstrate the potential for achieving compact and low-power chaotic oscillators through digital programmability and circuit integration.

3.6. Development directions and challenges

Intelligent systems are already being used in a wide range of applications, from virtual assistants and chatbots to self-driving cars and medical diagnoses. However, as these systems become more prevalent, they also face significant challenges, both in terms of technical limitations and ethical concerns. This section will explore the future of intelligent systems and the challenges they face.

3.6.1. Development directions

Intelligent systems have been advancing at a rapid pace, and they will continue to transform our lives in the coming years. There are some of the most promising directions:

- **Healthcare:** Intelligent systems can help diagnose diseases, monitor patient health, and provide personalized treatment recommendations. In addition, intelligent systems can also be used to develop new drugs and therapies.
- **Transportation:** Self-driving cars are already being tested on public roads, and they have the potential to improve road safety and reduce traffic congestion. Intelligent systems can also be used to optimize transportation routes, improve logistics, and reduce carbon emissions.

3.6.2. Challenges

Intelligent systems have the potential to transform our lives and revolutionize industries. However, they also face the following challenges:

- **Interpretability:** It is essential for intelligent systems to provide transparent and interpretable results, especially in critical decision-making processes. However, many of the state-of-the-art machine learning models are often considered "black-boxes," making it difficult to understand how they arrived at their results. This lack of interpretability can hinder trust in the system.
- **Cybersecurity and privacy:** Intelligent systems collect, store, and process a vast amount of data, which makes them vulnerable to cyber attacks. There is also a risk of data breaches that may compromise the privacy and security of individuals.

4. Optimization algorithms and strategies

Optimization is a fundamental process of finding the optimal solution within a given set of constraints. In computer science, optimization algorithms constitute a class of algorithms employed to obtain the optimal solution, and they can be categorized into two types:

- **Stochastic algorithms:** The stochastic algorithms leverage random properties to achieve better solutions through corresponding probabilistic strategies. Such algorithms fall into the category of optimization algorithms in computer science. Examples of commonly used stochastic algorithms include genetic algorithms, particle swarm algorithms, and beetle antennae search algorithms (Khan et al., 2022a). While these algorithms can find near-optimal solutions in a relatively short time, they are not guaranteed to obtain the optimal solution.
- **Deterministic algorithms:** The deterministic algorithms always generate the same output for a given input. Linear programming, integer programming, and dynamic programming are some examples of deterministic algorithms. These algorithms can provide efficient solutions to optimization problems. However, their computational power and time may be limited when dealing with complex optimization problems.

Subsequently, we will present an overview of bio-inspired optimization algorithms and intelligent optimization strategies.

4.1. Bio-inspired optimization algorithms

Bio-inspired optimization algorithms are a type of stochastic algorithms that draw inspiration from the principles of biological evolution and swarm intelligence observed in nature. These algorithms aim to mimic the behavior of individual organisms or groups for solving complex optimization problems (Khan et al., 2020b, 2022c; Chen et al., 2022b).

4.1.1. Particle swarm optimization (PSO) algorithm

In a study by Peng et al. (2020), an enhanced chaotic quantum-inspired particle swarm optimization (ICQPSO) algorithm was introduced to address the issues associated with Takagi–Sugeno fuzzy neural networks (TSFNNs), such as slow convergence rate and extended computation time. The flow chart illustrating the training and testing process of the ICQPSO algorithm for optimizing TSFNNs can be found in Figure 6. In another study by Yang et al. (2022), an improved particle swarm optimization (IPSO) algorithm was proposed to identify the parameters of the Preisach model, which is utilized to model hysteresis phenomena. The authors demonstrated that the IPSO algorithm outperformed the traditional PSO algorithm in terms of faster convergence, reduced computation time, and improved accuracy.

4.1.2. Genetic algorithm (GA)

In Ou et al. (2022), a hybrid knowledge extraction framework was developed by the authors, utilizing the combination of genetic algorithms and back propagation neural networks (BPNNs). An improved adaptive genetic algorithm (LAGA) was incorporated in the optimization of BPNNs. The efficacy of the LAGA-BPNNs approach was demonstrated through a case study involving the Wisconsin breast cancer dataset. Meanwhile, in Li et al. (2020a), the authors also investigated the applicability of the harmonic search algorithm to this knowledge extraction framework.

4.1.3. Cuckoo search (CS) algorithm

In Zhang et al. (2021), the authors presented an improved cuckoo search (ICS) algorithm that addressed the limitations of the original cuckoo search (CS) algorithm. The proposed ICS algorithm incorporated non-linear inertial weight, which enhances the local optimization capability, and the differential evolution algorithm, which improves convergence accuracy. The performance of the ICS algorithm was evaluated, and it was found to outperform the original CS algorithm in terms of both global search and robustness. In Ye et al. (2022), the authors

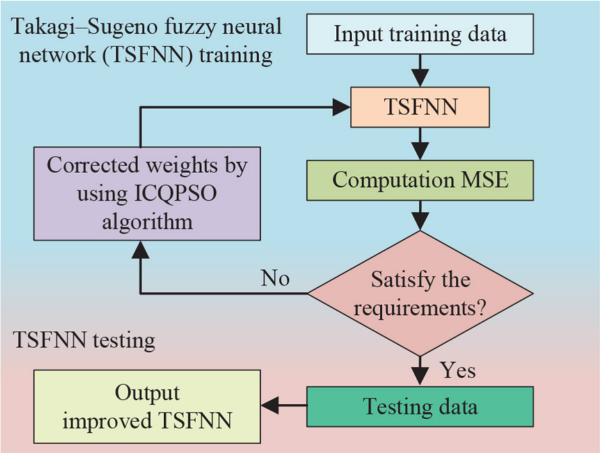

Takagi–Sugeno fuzzy neural network (TSFNN) training
Input training data
TSFNN
Corrected weights by using ICQPSO algorithm
Computation MSE
Satisfy the requirements?
No
Yes
TSFNN testing
Output improved TSFNN
Testing data

FIGURE 6
Training and testing flow chart for optimizing Takagi-Sugeno fuzzy neural networks (TSFNNs) by using an improved chaotic quantum particle swarm optimization (ICQPSO) algorithm. Mean square error (MSE) is a widely used metric to measure the average squared difference between the actual and predicted values of a regression problem. A lower MSE indicates that the predicted values are closer to the actual values, while a higher MSE indicates that the predictions are farther away from the actual values.

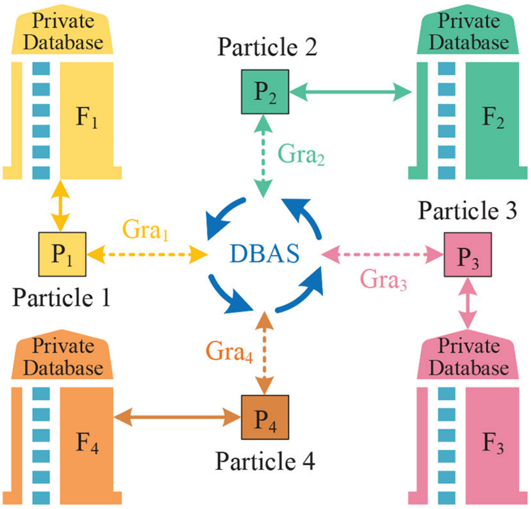


FIGURE 7
Framework of data exchange in the distributed beetle antenna search (DBAS) algorithm for solving multi-portfolio selection problem, where the search particles share only the gradient (Gra) and not the private information of the portfolio, such as customer information, stock information, and private databases.

proposed an improved multi-objective cuckoo search (IMOCS) algorithm to solve multi-objective optimization problems. The IMOCS algorithm demonstrated good convergence performance by dynamically adjusting the balance between development and exploration, compared to existing CS algorithms. The proposed algorithm provides an effective approach to deal with multi-objective optimization problems, which often involve multiple competing objectives.

4.1.4. Beetle antennae search (BAS) algorithm

In Khan et al. (2022d), a distributed beetle antennae search (DBAS) algorithm was proposed to solve the multi-portfolio selection problem, while ensuring privacy of investment portfolio data. The DBAS algorithm was shown to be efficient and robust in selecting the optimal investment portfolio. The paper also presented a data exchange framework for multi-portfolio selection, illustrated in Figure 7. In Liao et al. (2022c), the authors proposed a non-linearly activated beetle antenna search (NABAS) algorithm for fraud detection of publicly traded firms. They compared the performance of the NABAS algorithm to that of other popular methods, including the SVM-FK algorithm and the logistic regression model, and concluded that the proposed algorithm was more efficient and accurate for fraud detection. In Katsikis et al. (2021), a novel approach utilizing the BAS algorithm was proposed for solving the problem of time-varying mean-variance portfolio selection under transaction costs and cardinality constraints. This approach is based on state-of-the-art meta-heuristic optimization techniques and offers a more realistic solution to the problem as compared to conventional methods. The effectiveness of the proposed method was verified through numerical experiments and computer simulations, which demonstrated its superiority over traditional approaches. Overall, the study presents an online solution that addresses the limitations of static methods for solving time-varying financial problems.

4.2. Optimization strategies and systems

Optimization strategies and systems have become increasingly important across various fields as they offer effective solutions to complex problems by finding the best possible outcomes. In this subsection, we will provide an overview of the related research on optimization strategies and systems. Optimization strategies refer to the methods and techniques that are used to optimize a system or process. These strategies include but are not limited to heuristic algorithms, mathematical programming, and simulation-based optimization. Optimization systems, on the other hand, are computer programs or platforms that employ optimization strategies to solve complex problems. These systems can be standalone applications or integrated with other software tools. By exploring the latest research in optimization strategies and systems, we can gain a better understanding of how these techniques can be applied in different fields to improve efficiency, productivity, and overall performance.

4.2.1. Optimization strategies

In Chen et al. (2014), the authors presented a cooperative obstacle avoidance model and an improved obstacle avoidance (OA) algorithm for mobile wireless sensor networks, aimed at enhancing the adaptability and robustness of the network in complex environments. The proposed strategies optimized path planning and achieved higher obstacle avoidance efficiency by predicting the motion path of obstacles and defining the steering direction. In Xiang et al. (2021), the authors proposed a new approach for automatic skeleton design that utilizes physical simulation and optimization algorithms to better adapt to various application scenarios. The paper concludes that the proposed optimization strategy outperforms other mainstream optimizers in robot design and animation applications.

4.2.2. Optimization systems

The optimization system is a crucial tool to reduce the time and effort needed to find the optimal solution while guaranteeing its optimality. In Li and Zhang (2022), the authors presented an optimization system for generating benchmark dynamic test functions. The proposed system represents an advancement in the field of benchmark dynamic test functions, which is currently underdeveloped. In Deng et al. (2020), the authors proposed an optimal torque control system for controlling variable-speed wind turbines. As per the conclusion, this optimized system improved the effective wind speed estimation accuracy by 2%–7% and the efficiency of electrical energy generation by 0.35%. The proposed system offers a promising approach to enhancing the performance of wind turbines for electricity generation.

4.3. Development directions and challenges

Optimization algorithms and strategies have been widely used in various fields, including engineering, finance, and operations research, among others. The goal of optimization is to find the best solution to a problem within a given set of constraints. Optimization algorithms and strategies are continually evolving to meet the increasing demands of complex problems. This section will explore the future development and challenges of optimization algorithms and strategies.

4.3.1. Development directions

Optimization algorithms and strategies are constantly evolving, driven by advances in mathematics, computer science, and various application domains. There are some potential directions that optimization algorithms and strategies may be headed:

- **Deep learning-based optimization:** Deep learning techniques such as neural networks have shown tremendous success in various applications, including optimization. One potential direction is to use deep learning techniques to optimize the parameters of optimization algorithms, making them more efficient and effective.
- **Optimization with uncertainty:** Many real-world optimization problems involve uncertainty, such as noisy measurements, incomplete information, or uncertain parameters. One potential direction is to develop new optimization algorithms that can handle uncertainty explicitly, such as robust optimization or stochastic optimization.

4.3.2. Challenges

Despite the optimization algorithms and strategies have been widely developed and used, there are also significant challenges that need to be addressed:

- **Big data:** The growth of big data and the increasing complexity of data structures pose significant challenges for optimization algorithms and strategies. Dealing with large-scale, high-dimensional, and heterogeneous data requires advanced optimization techniques that can handle data efficiently and effectively.
- **Interdisciplinary applications:** Optimization problems are increasingly being used in interdisciplinary applications, such as healthcare, finance, energy, and transportation. These applications require optimization algorithms and strategies that can handle complex, multi-disciplinary problems, and that can effectively integrate domain knowledge, data analytics, and decision-making.

5. Conclusion

In this paper, we have analyzed and outlined the work related to neural networks, intelligent systems, and optimization algorithms and strategies in the rapidly evolving intelligence approach. Through an analysis and comparison of related work, we have shown that these intelligent approaches have rapidly evolved and have facilitated the efficient solution of practical problems. However, there are still emerging challenges that need to be addressed. Overall, this paper provides a valuable introduction and supplement to these important and rapidly evolving areas, highlighting their positive results and encouraging future research in these fields.

Author contributions

BL and SL developed the initial idea for the paper. CH and BL conducted the literature review and analyzed the relevant studies. CH wrote the first draft of the paper. CH and XC reviewed and edited the paper. XC, BL, and SL provided supervision and support for the entire project and offering guidance and assistance throughout the writing process. BL secured funding for the paper. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62066015, the Natural Science Foundation of Hunan Province of China under Grant No. 2020JJ4511, and the Research Foundation of Education Bureau of Hunan Province, China, under Grant No. 20A396.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Chen, C., Zain, A. M., and Zhou, K. (2022a). Definition, approaches, and analysis of code duplication detection (2006–2020): a critical review. *Neural Comput. Appl.* 34, 20507–05371. doi: 10.1007/s00521-022-07707-2

Chen, S., Jin, J., and Zhang, W. (2013). Low-voltage low-power folded mixer using current-reuse technical for IEEE 802.11 b wireless application. *IETE J. Res.* 59, 415–419. doi: 10.4103/0377-2063.118066

Chen, Z., Ding, L., Chen, K., and Li, R. (2014). The study of cooperative obstacle avoidance method for MWSN based on flocking control. *Sci. World J.* 2014, 614346. doi: 10.1155/2014/614346

Chen, Z., Francis, A., Li, S., Liao, B., Xiao, D., Ha, T. T., et al. (2022b). Egret swarm optimization algorithm: an evolutionary computation approach for model free optimization. *Biomimetics* 7, 144. doi: 10.3390/biomimetics7040144

Dai, H., Tan, W., and Zheng, Z. (2018). Spatio-temporal dynamics and interaction of lump solutions for the (4+ 1)-d fokas equation. *Thermal Sci.* 22, 1823–1830. doi: 10.2298/TSCI1804823D

Dai, J., Chen, Y., Xiao, L., Jia, L., and He, Y. (2022). Design and analysis of a hybrid GNN-ZNN model with a fuzzy adaptive factor for matrix inversion. *IEEE Trans. Indus. Inform.* 18, 2434–2442. doi: 10.1109/TII.2021.3093115

Deng, X., Yang, J., Sun, Y., Song, D., Xiang, X., Ge, X., et al. (2019). Sensorless effective wind speed estimation method based on unknown input disturbance observer and extreme learning machine. *Energy* 186, 115790. doi: 10.1016/j.energy.2019.07.120

Deng, X., Yang, J., Sun, Y., Song, D., Yang, Y., and Joo, Y. H. (2020). An effective wind speed estimation based extended optimal torque control for maximum wind energy capture. *IEEE Access* 8, 65959–65969. doi: 10.1109/ACCESS.2020.2984654

Ding, B., Li, X., and Li, Y. (2022). Configuration design and experimental verification of a variable constant-force compliant mechanism. *Robotica* 40, 3463–3475. doi: 10.1017/S0263574722000340

Ding, B., Yang, Z., Xiao, X., and Zhang, G. (2019a). Design of reconfigurable planar micro-positioning stages based on function modules. *IEEE Access* 7, 15102–15112. doi: 10.1109/ACCESS.2019.2894619

Ding, B., Zhao, J., and Li, Y. (2021). Design of a spatial constant-force end-effector for polishing/deburring operations. *Int. J. Adv. Manufact. Technol.* 116, 3507–3515. doi: 10.1007/s00170-021-07579-1

Ding, L., She, J., and Peng, S. (2013). An integrated prediction model for network traffic based on wavelet transformation. *Elektronika ir elektrotechnika* 19, 73–76. doi: 10.5755/j01.eee.19.3.3700

Ding, L., Xiao, L., Zhou, K., Lan, Y., and Zhang, Y. (2018). A new RNN model with a modified nonlinear activation function applied to complex-valued linear equations. *IEEE Access* 6, 62954–62962. doi: 10.1109/ACCESS.2018.2876665

Ding, L., Xiao, L., Zhou, K., Lan, Y., Zhang, Y., and Li, J. (2019b). An improved complex-valued recurrent neural network model for time-varying complex-valued Sylvester equation. *IEEE Access* 7, 19291–19302. doi: 10.1109/ACCESS.2019.2896983

Ding, L., Zeng, H., Wang, W., and Yu, F. (2014). Improved stability criteria of static recurrent neural networks with a time-varying delay. *Sci. World J.* 2014, 391282. doi: 10.1155/2014/391282

Guo, D., Li, K., and Liao, B. (2018). Bi-criteria minimization with MWVN–INAM type for motion planning and control of redundant robot manipulators. *Robotica* 36, 655–675. doi: 10.1017/S0263574717000625

Han, L., He, Y., Liao, B., and Hua, C. (2023). An accelerated double-integral ZNN with resisting linear noise for dynamic Sylvester equation solving and its application to the control of the SFM chaotic system. *Axioms* 12, 287. doi: 10.3390/axioms12030287

Hu, H., Fang, M., Hu, F., Zeng, S., and Deng, X. (2021). A new design of substation grounding based on electrolytic cathodic protection and on transfer corrosion current. *Electric Power Syst. Res.* 195, 107174. doi: 10.1016/j.epsr.2021.107174

Hu, H., Luo, R., Fang, M., Zeng, S., and Hu, F. (2019a). A new optimization design for grounding grid. *Int. J. Electrical Power Energy Syst.* 108, 61–71. doi: 10.1016/j.ijepes.2018.12.041

Hu, K., He, W., Ye, J., Zhao, L., Peng, H., and Pi, J. (2019b). Online visual tracking of weighted multiple instance learning via neutrosophic similarity-based objectness estimation. *Symmetry* 11, 832. doi: 10.3390/sym11060832

Huang, Z., Dong, M., Du, X., and Guan, Y. (2022). A nonlinear prediction model of antislide pile top displacement based on MIC-SVR for jurassic landslides. *Adv. Civil Eng.* 2022, 9101234. doi: 10.1155/2022/9101234

Jia, L., Xiao, L., Dai, J., Qi, Z., Zhang, Z., and Zhang, Y. (2021). Design and application of an adaptive fuzzy control strategy to zeroing neural network for solving time-variant QP problem. *IEEE Trans. Fuzzy Syst.* 29, 1544–1555. doi: 10.1109/TFUZZ.2020.2981001

Jian, Z., Xiao, L., Li, K., Zuo, Q., and Zhang, Y. (2020). Adaptive coefficient designs for nonlinear activation function and its application to zeroing neural network for solving time-varying Sylvester equation. *J. Franklin Instit.* 357, 9909–9929. doi: 10.1016/j.jfranklin.2020.06.029

Jin, J. (2014). Resistorless active simo universal filter and four-phase quadrature oscillator. *Arab. J. Sci. Eng.* 39, 3887–3894. doi: 10.1007/s13369-014-0985-y

Jin, J. (2018a). Novel quadrature voltage-controlled oscillator using capacitor coupling. *IETE J. Res.* 64, 263–269. doi: 10.1080/03772063.2017.1351318

Jin, J. (2018b). Programmable multi-direction fully integrated chaotic oscillator. *Microelectron. J.* 75, 27–34. doi: 10.1016/j.mejo.2018.02.007

Jin, J., and Fu, K. (2013). An ultra-low-power integrated rf receiver for multi-standard wireless applications. *IETE J. Res.* 59, 447–453. doi: 10.4103/0377-2063.118022

Jin, J., and Liang, P. (2013). Resistorless current-mode quadrature oscillator with grounded capacitors. *Rev. Roum. Sci. Technol.–Électrotechn. Énerg* 58, 304–313.

Jin, J., and Tan, M. (2019). Low power quadrature voltage controlled oscillator. *Int. J. RF Microwave Comput. Aided Eng.* 29, e21952. doi: 10.1002/mmce.21952

Jin, J., Wang, C., and Sun, J. (2015a). Novel third-order quadrature oscillators with grounded capacitors. *Automatika* 56, 207–216. doi: 10.7305/automatika.2015.07.669

Jin, J., Wang, C., Sun, J., and Du, S. (2015b). Design and simulation of novel amplifier-based mixer for ISM band wireless applications. *Int. J. Circ. Theory Appl.* 43, 1794–1800. doi: 10.1002/cta.2028

Jin, J., Wang, C., Sun, J., Tu, Y., Zhao, L., and Xia, Z. (2014a). Novel digitally programmable multiphase voltage controlled oscillator and its stability discussion. *Microelectron. Reliabil.* 54, 595–600. doi: 10.1016/j.microrel.2013.12.008

Jin, J., Wang, C., Xia, Z., and Yang, H. (2014b). Sub-harmonic upconversion mixer using 0.18 μm cmos technology. *Electron. Lett.* 50, 1955–1957. doi: 10.1049/el.2014.3026

Jin, J., Xiao, L., Lu, M., and Li, J. (2019). Design and analysis of two FTRNN models with application to time-varying Sylvester equation. *IEEE Access* 7, 58945–58950. doi: 10.1109/ACCESS.2019.2911130

Jin, J., and Yu, F. (2012). Arming antennas with dual bandstops. *Microwaves RF* 51, 56.

Jin, J., and Yu, F. (2013). Novel current-reuse current-mirror and its application on 2.4-GHz down-conversion mixer. *Microwave Opt. Technol. Lett.* 55, 2520–2524. doi: 10.1002/mop.27878

Jin, J., Zhou, K., and Zhao, L. (2017a). Designing rf ring oscillator using current-mode technology. *IEEE Access* 5, 5306–5312. doi: 10.1109/ACCESS.2017.2692771

Jin, L., Li, S., Xiao, L., Lu, R., and Liao, B. (2017b). Cooperative motion generation in a distributed network of redundant robot manipulators with noises. *IEEE Trans. Syst. Man Cybernet. Syst.* 48, 1715–1724. doi: 10.1109/TSMC.2017.2693400

Katsikis, V. N., and Mourtas, S. D. (2021). "Portfolio insurance and intelligent algorithms," in *Computational Management: Applications of Computational Intelligence in Business Management* (Springer), 305–323. doi: 10.1007/978-3-030-72929-5_14

Katsikis, V. N., Mourtas, S. D., Stanimirović, P. S., Li, S., and Cao, X. (2021). "Time-varying mean-variance portfolio selection under transaction costs and cardinality constraint problem via beetle antennae search algorithm (BAS)," in *Operations Research Forum*, Vol. 2 (Springer), 1–26. doi: 10.1007/s43069-021-00060-5

Katsikis, V. N., Stanimirović, P. S., Mourtas, S. D., Xiao, L., Stanujkić, D., and Karabašević, D. (2023). Zeroing neural network based on neutrosophic logic for calculating minimal-norm least-squares solutions to time-varying linear systems. *Neural Process. Lett.* doi: 10.1109/TNNLS.2022.3171715-7

Khan, A. H., Cao, X., Li, S., Katsikis, V. N., and Liao, L. (2020a). BAS-Adam: an Adam based approach to improve the performance of beetle antennae search optimizer. *IEEE/CAA J. Automat. Sin.* 7, 461–471. doi: 10.1109/JAS.2020.1003048

Khan, A. H., Cao, X., Li, S., and Luo, C. (2020b). Using social behavior of beetles to establish a computational model for operational management. *IEEE Trans. Comput. Soc. Syst.* 7, 492–502. doi: 10.1109/TCSS.2019.2958522

Khan, A. H., Li, S., and Cao, X. (2021). Tracking control of redundant manipulator under active remote center-of-motion constraints: an RNN-based metaheuristic approach. *Sci. China Inform. Sci.* 64, 1–18. doi: 10.1007/s11432-019-2735-6

Khan, A. T., Cao, X., Brajevic, I., Stanimirovic, P. S., Katsikis, V. N., and Li, S. (2022a). Non-linear activated beetle antennae search: a novel technique for non-convex tax-aware portfolio optimization problem. *Expert Syst. Appl.* 197, 116631. doi: 10.1016/j.eswa.2022.116631

Khan, A. T., Cao, X., and Li, S. (2022b). Dual beetle antennae search system for optimal planning and robust control of 5-link biped robots. *J. Comput. Sci.* 60, 101556. doi: 10.1016/j.jocs.2022.101556

Khan, A. T., Cao, X., Li, S., Katsikis, V. N., Brajevic, I., and Stanimirovic, P. S. (2022c). Fraud detection in publicly traded us firms using beetle antennae search: a machine learning approach. *Expert Syst. Appl.* 191, 116148. doi: 10.1016/j.eswa.2021.116148

Khan, A. T., Cao, X., Liao, B., and Francis, A. (2022d). Bio-inspired machine learning for distributed confidential multi-portfolio selection problem. *Biomimetics* 7, 124. doi: 10.3390/biomimetics7030124

Khan, A. T., Li, S., and Cao, X. (2022e). Human guided cooperative robotic agents in smart home using beetle antennae search. *Sci. China Inform. Sci.* 65, 122204. doi: 10.1007/s11432-020-3073-5

Lan, Y., Wang, L., Chen, C., and Ding, L. (2017). Optimal sliding mode robust control for fractional-order systems with application to permanent magnet synchronous motor tracking control. *J. Optimizat. Theory Appl.* 174, 197–209. doi: 10.1007/s10957-015-0827-4

Lan, Y., Wang, L., Ding, L., and Zhou, Y. (2016). Full-order and reduced-order observer design for a class of fractional-order nonlinear systems. *Asian J. Control* 18, 1467–1477. doi: 10.1002/asjc.1230

Lei, K., Yang, X., Tan, Y., Peng, S., and Cao, X. (2016). Principal component analysis-based blind wideband spectrum sensing for cognitive radio. *Electron. Lett.* 52, 1416–1418. doi: 10.1049/el.2016.1319

Lei, Y., Liao, B., and Chen, J. (2020). Comprehensive analysis of ZNN models for computing complex-valued time-dependent matrix inverse. *IEEE Access* 8, 91989–91998. doi: 10.1109/ACCESS.2020.2994102

Lei, Y., Liao, B., and Yin, Q. (2019). A noise-acceptable ZNN for computing complex-valued time-dependent matrix pseudoinverse. *IEEE Access* 7, 13832–13841. doi: 10.1109/ACCESS.2019.2894180

Lei, Y., Luo, J., Chen, T., Ding, L., Liao, B., Xia, G., et al. (2022). Nonlinearly activated IEZNN model for solving time-varying Sylvester equation. *IEEE Access* 10, 121520–121530. doi: 10.1109/ACCESS.2022.3222372

Li, H., and Zhang, G. (2022). Designing benchmark generator for dynamic optimization algorithm. *IEEE Access* 10, 638–648. doi: 10.1109/ACCESS.2021.3138141

Li, H., Zhou, K., Mo, L., Zain, A. M., and Qin, F. (2020a). Weighted fuzzy production rule extraction using modified harmony search algorithm and bp neural network framework. *IEEE Access* 8, 186620–186637. doi: 10.1109/ACCESS.2020.3029966

Li, J., Peng, H., Hu, H., Luo, Z., and Tang, C. (2020b). Multimodal information fusion for automatic aesthetics evaluation of robotic dance poses. *Int. J. Soc. Robot.* 12, 5–20. doi: 10.1007/s12369-019-00535-w

Li, M., Dai, H., Wei, X., and Tan, W. (2022a). Some new soliton solutions and dynamical behaviours of (3+ 1)-dimensional Jimbo-Miwa equation. *Int. J. Comput. Math.* 99, 1654–1668. doi: 10.1080/00207160.2021.1998468

Li, S., and Li, Y. (2013). Nonlinearly activated neural network for solving time-varying complex Sylvester equation. *IEEE Trans. Cybernet.* 44, 1397–1407. doi: 10.1109/TCYB.2013.2285166

Li, W., Han, L., Xiao, X., Liao, B., and Peng, C. (2022b). A gradient-based neural network accelerated for vision-based control of a RCM-constrained surgical endoscope robot. *Neural Comput. Appl.* 34, 1329–1343. doi: 10.1007/s00521-021-06465-x

Li, W., Xiao, L., and Liao, B. (2020c). A finite-time convergent and noise-rejection recurrent neural network and its discretization for dynamic nonlinear equations solving. *IEEE Trans. Cybernet.* 50, 3195–3207. doi: 10.1109/TCYB.2019.2906263

Li, Z., Liao, B., Xu, F., and Guo, D. (2020d). A new repetitive motion planning scheme with noise suppression capability for redundant robot manipulators. *IEEE Trans. Syst. Man Cybernet. Syst.* 50, 5244–5254. doi: 10.1109/TSMC.2018.2870523

Liao, B., Han, L., Cao, X., Li, S., and Li, J. (2023). Double integral-enhanced zeroing neural network with linear noise rejection for time-varying matrix inverse. *CAAI Trans. Intell. Technol.* doi: 10.1049/cit2.12161

Liao, B., Han, L., He, Y., Cao, X., and Li, J. (2022a). Prescribed-time convergent adaptive ZNN for time-varying matrix inversion under harmonic noise. *Electronics* 11, 1636. doi: 10.3390/electronics11101636

Liao, B., Hua, C., Cao, X., Katsikis, V. N., and Li, S. (2022b). Complex noise-resistant zeroing neural network for computing complex time-dependent Lyapunov equation. *Mathematics* 10, 2817. doi: 10.3390/math10152817

Liao, B., Huang, Z., Cao, X., and Li, J. (2022c). Adopting nonlinear activated beetle antennae search algorithm for fraud detection of public trading companies: a computational finance approach. *Mathematics* 10, 2160. doi: 10.3390/math10132160

Liao, B., and Liu, W. (2015). Pseudoinverse-type bi-criteria minimization scheme for redundancy resolution of robot manipulators. *Robotica* 33, 2100–2113. doi: 10.1017/S0263574714001349

Liao, B., Wang, Y., Li, J., Guo, D., and He, Y. (2022d). Harmonic noise-tolerant ZNN for dynamic matrix pseudoinversion and its application to robot manipulator. *Front. Neurorobot.* 16, 928636. doi: 10.3389/fnbot.2022.928636

Liao, B., Wang, Y., Li, W., Peng, C., and Xiang, Q. (2021). Prescribed-time convergent and noise-tolerant Z-type neural dynamics for calculating time-dependent quadratic programming. *Neural Comput. Appl.* 33, 5327–5337. doi: 10.1007/s00521-020-05356-x

Liao, B., Xiang, Q., and Li, S. (2019). Bounded Z-type neurodynamics with limited-time convergence and noise tolerance for calculating time-dependent Lyapunov equation. *Neurocomputing* 325, 234–241. doi: 10.1016/j.neucom.2018.10.031

Liao, B., Zhang, Y., and Jin, L. (2015). Taylor $O(h^3)$ discretization of znn models for dynamic equality-constrained quadratic programming with application to manipulators. *IEEE Trans. Neural Netw. Learn. Syst.* 27, 225–237. doi: 10.1109/TNNLS.2015.2435014

Liao, S., Ding, B., and Li, Y. (2022e). Design, assembly, and simulation of flexure-based modular micro-positioning stages. *Machines* 10, 421. doi: 10.3390/machines10060421

Liu, M., Liao, B., Ding, L., and Xiao, L. (2016). Performance analyses of recurrent neural network models exploited for online time-varying nonlinear optimization. *Comput. Sci. Inform. Syst.* 13, 691–705. doi: 10.2298/CSIS160215023L

Long, F., Ding, L., and Li, J. (2022). DGFlow-SLAM: a novel dynamic environment RGB-D SLAM without prior semantic knowledge based on grid segmentation of scene flow. *Biomimetics* 7, 163. doi: 10.3390/biomimetics7040163

Lu, H., Jin, L., Luo, X., Liao, B., Guo, D., and Xiao, L. (2019). RNN for solving perturbed time-varying underdetermined linear system with double bound limits on residual errors and state variables. *IEEE Trans. Indus. Inform.* 15, 5931–5942. doi: 10.1109/TII.2019.2909142

Lu, S., Li, Y., and Ding, B. (2020). Kinematics and dynamics analysis of the 3PUS-PRU parallel mechanism module designed for a novel 6-DOF gantry hybrid machine tool. *J. Mech. Sci. Technol.* 34, 345–357. doi: 10.1007/s12206-019-1234-9

Luo, W., Ou, Q., Yu, F., Cui, L., and Jin, J. (2020). Analysis of a new hidden attractor coupled chaotic system and application of its weak signal detection. *Math. Prob. Eng.* 2020, 1–15. doi: 10.1155/2020/8849283

Luo, Z., and Xie, J. (2017). Multiple periodic solutions for a class of second-order neutral functional differential equations. *Adv. Diff. Equat.* 2017, 1–8. doi: 10.1186/s13662-016-1064-3

Lv, X., Xiao, L., Tan, Z., and Yang, Z. (2018). WSBP function activated Zhang dynamic with finite-time convergence applied to Lyapunov equation. *Neurocomputing* 314, 310–315. doi: 10.1016/j.neucom.2018.06.057

Niu, Y., Peng, C., and Liao, B. (2022). Batch-wise permutation feature importance evaluation and problem-specific bigraph for learn-to-branch. *Electronics* 11, 2253. doi: 10.3390/electronics11142253

Ou, Y., Ye, S., Ding, L., Zhou, K., and Zain, A. M. (2022). Hybrid knowledge extraction framework using modified adaptive genetic algorithm and BPNN. *IEEE Access* 10, 72037–72050. doi: 10.1109/ACCESS.2022.3188689

Ouyang, Z., Jin, J., Yu, F., Chen, L., and Ding, L. (2022). Fully integrated chen chaotic oscillation system. *Discrete Dyn. Nat. Soc.* 2022, 8613090. doi: 10.1155/2022/8613090

Peng, C., and Liao, B. (2022). Heavy-head sampling for fast imitation learning of machine learning based combinatorial auction solver. *Neural Process. Lett.* 55, 631–644. doi: 10.1007/s11063-022-10900-y

Peng, H., Hu, H., Chao, F., Zhou, C., and Li, J. (2016). Autonomous robotic choreography creation via semi-interactive evolutionary computation. *Int. J. Soc. Robot.* 8, 649–661. doi: 10.1007/s12369-016-0355-x

Peng, H., Li, J., Hu, H., Hu, K., Tang, C., and Ding, Y. (2019a). Creating a computable cognitive model of visual aesthetics for automatic aesthetics evaluation of robotic dance poses. *Symmetry* 12, 23. doi: 10.3390/sym12010023

Peng, H., Li, J., Hu, H., Hu, K., Zhao, L., and Tang, C. (2022). Automatic aesthetics assessment of robotic dance motions. *Robot. Auton. Syst.* 155, 104160. doi: 10.1016/j.robot.2022.104160

Peng, H., Li, J., Hu, H., Zhao, L., Feng, S., and Hu, K. (2019b). Feature fusion based automatic aesthetics evaluation of robotic dance poses. *Robot. Auton. Syst.* 111, 99–109. doi: 10.1016/j.robot.2018.10.016

Peng, H., Zhou, C., Hu, H., Chao, F., and Li, J. (2015). Robotic dance in social robotics—A taxonomy. *IEEE Trans. Hum. Mach. Syst.* 45, 281–293. doi: 10.1109/THMS.2015.2393558

Peng, Y., Lei, K., Yang, X., and Peng, J. (2020). Improved chaotic quantum-behaved particle swarm optimization algorithm for fuzzy neural network and its application. *Math. Prob. Eng.* 2020, 1–11. doi: 10.1155/2020/9464593

Qin, H., Wang, C., Xi, X., Tian, J., and Zhou, G. (2017). Simulating the effects of the airborne lidar scanning angle, flying altitude, and pulse density for forest foliage profile retrieval. *Appl. Sci.* 7, 712. doi: 10.3390/app7070712

Qu, C., He, W., Peng, X., and Peng, X. (2020). Harris Hawks optimization with information exchange. *Appl. Math. Model.* 84, 52–75. doi: 10.1016/j.apm.2020.03.024

Song, D., Yang, Y., Zheng, S., Deng, X., Yang, J., Su, M., et al. (2020). New perspectives on maximum wind energy extraction of variable-speed wind turbines using previewed wind speeds. *Energy Conv. Manage.* 206, 112496. doi: 10.1016/j.enconman.2020.112496

Sun, H., Zhang, Y., and Yang, X. (2022). 6.25-10Gb/s adaptive CTLE with spectrum balancing and loop-unrolled half-rate DFE in TSMC 0.18 μm CMOS. *IEICE Electron. Express* 19, 20220429. doi: 10.1587/elex.19.20220429

Sun, Y., Yu, Z., and Wang, Z. (2016). Bioinspired design of building materials for blast and ballistic protection. *Adv. Civil Eng.* 2016, 5840176. doi: 10.1155/2016/5840176

Tan, W. (2021). Some new dynamical behaviour of double breathers and lump-n-solitons for the ITO equation. *Int. J. Comput. Math.* 98, 961–974. doi: 10.1080/00207160.2020.1792454

Tan, W., and Dai, Z. (2016). Dynamics of kinky wave for (3+ 1)-dimensional potential Yu–Toda–Sasa–Fukuyama equation. *Nonlinear Dyn.* 85, 817–823. doi: 10.1007/s11071-016-2725-1

Tan, W., and Dai, Z. (2017). Spatiotemporal dynamics of lump solution to the (1+ 1)-dimensional benjamin–ono equation. *Nonlinear Dyn.* 89, 2723–2728. doi: 10.1007/s11071-017-3620-0

Tan, W., Dai, Z., and Dai, H. (2017). Dynamical analysis of lump solution for the (2+ 1)-dimensional ITO equation. *Thermal Sci.* 21, 1673–1679. doi: 10.2298/TSCI160812145T

Tan, W., Dai, Z.-D., and Yin, Z. (2019a). Dynamics of multi-breathers, n-solitons and m-lump solutions in the (2+ 1)-dimensional KDV equation. *Nonlinear Dyn.* 96, 1605–1614. doi: 10.1007/s11071-019-04873-2

Tan, Z., Hu, Y., Xiao, L., and Chen, K. (2019b). Robustness analysis and robotic application of combined function activated RNN for time-varying matrix pseudo inversion. *IEEE Access* 7, 33434–33440. doi: 10.1109/ACCESS.2019.2904605

Tang, A.-Y., Li, X.-F., Wu, J.-X., and Lee, K. (2015). Flapwise bending vibration of rotating tapered Rayleigh cantilever beams. *J. Construct. Steel Res.* 112, 1–9. doi: 10.1016/j.jcsr.2015.04.010

Tang, C., Hu, H., Wang, W., Li, W., Peng, H., and Wang, X. (2020). Using a multilearner to fuse multimodal features for human action recognition. *Math. Prob. Eng.* 2020, 1–18. doi: 10.1155/2020/5892312

Tang, Z., Tan, N., and Zhang, Y. (2022). Velocity-layer Zhang equivalency for time-varying joint limits avoidance of redundant robot manipulator. *IET Control Theory Appl.* 16, 1909–1921. doi: 10.1049/cth2.12355

Wu, C., Hao, T., Qi, L., Zhuo, D., Feng, Z., Zhang, J., et al. (2022a). Multi-feature extraction-based defect recognition of foundation pile under layered soil condition using convolutional neural network. *Appl. Sci.* 12, 9840. doi: 10.3390/app12199840

Wu, C., Peng, Y., Zhuo, D., Zhang, J., Ren, W., and Feng, Z. (2022b). Energy ratio variation-based structural damage detection using convolutional neural network. *Appl. Sci.* 12, 10220. doi: 10.3390/app122010220

Wu, C., Zhang, J., Qi, L., and Zhuo, D. (2022c). Defect identification of concrete piles based on numerical simulation and convolutional neural network. *Buildings* 12, 664. doi: 10.3390/buildings12050664

Wu, S., Yan, H., Wang, Z., Bi, R., and Jia, L. (2021). Tool profile modification of hypoid gear machined by the duplex helical method. *Int. J. Adv. Manufact. Technol.* 119, 3771–3784. doi: 10.1007/s00170-021-08461-w

Xiang, C., Zhang, K., Jiang, C., Long, D., He, Q., Zhou, X., et al. (2022). A scheme to restrain PAPR and frequency selective fading in 64 QAM MB-OFDM UWBoF system. *Opt. Fiber Technol.* 73, 103040. doi: 10.1016/j.yofte.2022.103040

Xiang, Q., Li, W., Liao, B., and Huang, Z. (2018a). Noise-resistant discrete-time neural dynamics for computing time-dependent Lyapunov equation. *IEEE Access* 6, 45359–45371. doi: 10.1109/ACCESS.2018.2863736

Xiang, X., Zhang, X., and Mo, X. (2018b). Statistical identification of Markov chain on trees. *Math. Prob. Eng.* 2018, 2036248. doi: 10.1155/2018/2036248

Xiang, Z., Xiang, C., Li, T., and Guo, Y. (2021). A self-adapting hierarchical actions and structures joint optimization framework for automatic design of robotic and animation skeletons. *Soft Comput.* 25, 263–276. doi: 10.1007/s00500-020-05139-5

Xiao, L. (2015). A finite-time convergent neural dynamics for online solution of time-varying linear complex matrix equation. *Neurocomputing* 167, 254–259. doi: 10.1016/j.neucom.2015.04.070

Xiao, L. (2016). A nonlinearly-activated neurodynamic model and its finite-time solution to equality-constrained quadratic optimization with nonstationary coefficients. *Appl. Soft Comput.* 40, 252–259. doi: 10.1016/j.asoc.2015.11.023

Xiao, L. (2017a). Accelerating a recurrent neural network to finite-time convergence using a new design formula and its application to time-varying matrix square root. *J. Franklin Instit.* 354, 5667–5677. doi: 10.1016/j.jfranklin.2017.06.012

Xiao, L. (2017b). A finite-time recurrent neural network for solving online time-varying Sylvester matrix equation based on a new evolution formula. *Nonlinear Dyn.* 90, 1581–1591. doi: 10.1007/s11071-017-3750-4

Xiao, L. (2019). A finite-time convergent Zhang neural network and its application to real-time matrix square root finding. *Neural Comput. Appl.* 31, 793–800. doi: 10.1007/s00521-017-3010-z

Xiao, L., Dai, J., Jin, L., Li, W., Li, S., and Hou, J. (2019a). A noise-enduring and finite-time zeroing neural network for equality-constrained time-varying nonlinear optimization. *IEEE Trans. Syst. Man Cybernet. Syst.* 51, 4729–4740. doi: 10.1109/TSMC.2019.2944152

Xiao, L., and He, Y. (2021). A noise-suppression ZNN model with new variable parameter for dynamic Sylvester equation. *IEEE Trans. Indus. Inform.* 17, 7513–7522. doi: 10.1109/TII.2021.3058343

Xiao, L., He, Y., and Liao, B. (2022a). A parameter-changing zeroing neural network for solving linear equations with superior fixed-time convergence. *Expert Syst. Appl.* 208, 118086. doi: 10.1016/j.eswa.2022.118086

Xiao, L., Jia, L., Zhang, Y., Hu, Z., and Dai, J. (2019b). Finite-time convergence and robustness analysis of two nonlinear activated ZNN models for time-varying linear matrix equations. *IEEE Access* 7, 135133–135144. doi: 10.1109/ACCESS.2019.2941961

Xiao, L., Li, K., and Duan, M. (2019c). Computing time-varying quadratic optimization with finite-time convergence and noise tolerance: a unified framework for zeroing neural network. *IEEE Trans. Neural Netw. Learn. Syst.* 30, 3360–3369. doi: 10.1109/TNNLS.2019.2891252

Xiao, L., Li, S., Li, K., Jin, L., and Liao, B. (2020a). Co-design of finite-time convergence and noise suppression: a unified neural model for time varying linear equations with robotic applications. *IEEE Trans. Syst. Man Cybernet. Syst.* 50, 5233–5243. doi: 10.1109/TSMC.2018.2870489

Xiao, L., Li, S., Lin, F., Tan, Z., and Khan, A. H. (2018a). Zeroing neural dynamics for control design: comprehensive analysis on stability, robustness, and convergence speed. *IEEE Trans. Indus. Inform.* 15, 2605–2616. doi: 10.1109/TII.2018.2867169

Xiao, L., Liao, B., Jin, J., Lu, R., Yang, X., and Ding, L. (2017a). A finite-time convergent dynamic system for solving online simultaneous linear equations. *Int. J. Comput. Math.* 94, 1778–1786. doi: 10.1080/00207160.2016.1247436

Xiao, L., Liao, B., Li, S., and Chen, K. (2018b). Nonlinear recurrent neural networks for finite-time solution of general time-varying linear matrix equations. *Neural Netw.* 98, 102–113. doi: 10.1016/j.neunet.2017.11.011

Xiao, L., Liao, B., Li, S., Zhang, Z., Ding, L., and Jin, L. (2017b). Design and analysis of FTZNN applied to the real-time solution of a nonstationary Lyapunov equation and tracking control of a wheeled mobile manipulator. *IEEE Trans. Indus. Inform.* 14, 98–105. doi: 10.1109/TII.2017.2717020

Xiao, L., Liu, S., Wang, X., He, Y., Jia, L., and Xu, Y. (2021a). Zeroing neural networks for dynamic quaternion-valued matrix inversion. *IEEE Trans. Indus. Inform.* 18, 1562–1571. doi: 10.1109/TII.2021.3090063

Xiao, L., and Lu, R. (2017). A fully complex-valued gradient neural network for rapidly computing complex-valued linear matrix equations. *Chinese J. Electron.* 26, 1194–1197. doi: 10.1049/cje.2017.06.007

Xiao, L., and Lu, R. (2019). A finite-time recurrent neural network for computing quadratic minimization with time-varying coefficients. *Chinese J. Electron.* 28, 253–258. doi: 10.1049/cje.2019.01.009

Xiao, L., Tao, J., Dai, J., Wang, Y., Jia, L., and He, Y. (2021b). A parameter-changing and complex-valued zeroing neural-network for finding solution of time-varying complex linear matrix equations in finite time. *IEEE Trans. Indus. Inform.* 17, 6634–6643. doi: 10.1109/TII.2021.3049413

Xiao, L., Tao, J., and Li, W. (2022b). An arctan-type varying-parameter ZNN for solving time-varying complex Sylvester equations in finite time. *IEEE Trans. Indus. Inform.* 18, 3651–3660. doi: 10.1109/TII.2021.3111816

Xiao, L., Yi, Q., Zuo, Q., and He, Y. (2020b). Improved finite-time zeroing neural networks for time-varying complex Sylvester equation solving. *Math. Comput. Simul.* 178, 246–258. doi: 10.1016/j.matcom.2020.06.014

Xiao, L., and Zhang, Y. (2014). Solving time-varying inverse kinematics problem of wheeled mobile manipulators using Zhang neural network with exponential convergence. *Nonlinear Dyn.* 76, 1543–1559. doi: 10.1007/s11071-013-1227-7

Xiao, L., Zhang, Y., Dai, J., Li, J., and Li, W. (2019d). New noise-tolerant ZNN models with predefined-time convergence for time-variant Sylvester equation solving. *IEEE Trans. Syst. Man. Cybernet. Syst.* 51, 3629–3640. doi: 10.1109/TSMC.2019.2930646

Xiao, L., Zhang, Y., Dai, J., Zuo, Q., and Wang, S. (2020c). Comprehensive analysis of a new varying parameter zeroing neural network for time varying matrix inversion. *IEEE Trans. Indus. Inform.* 17, 1604–1613. doi: 10.1109/TII.2020.2989173

Xiao, L., Zhang, Y., Hu, Z., and Dai, J. (2019e). Performance benefits of robust nonlinear zeroing neural network for finding accurate solution of Lyapunov equation in presence of various noises. *IEEE Trans. Indus. Inform.* 15, 5161–5171. doi: 10.1109/TII.2019.2900659

Xiao, L., Zhang, Y., Zuo, Q., Dai, J., Li, J., and Tang, W. (2019f). A noise-tolerant zeroing neural network for time-dependent complex matrix inversion under various kinds of noises. *IEEE Trans. Indus. Inform.* 16, 3757–3766. doi: 10.1109/TII.2019.2936877

Xiao, L., Zhang, Z., and Li, S. (2019g). Solving time-varying system of nonlinear equations by finite-time recurrent neural networks with application to motion tracking of robot manipulators. *IEEE Trans. Syst. Man Cybernet. Syst.* 49, 2210–2220. doi: 10.1109/TSMC.2018.2836968

Xu, X., Zhang, R., and Qian, Y. (2022). Location-based hybrid precoding schemes and QOS-aware power allocation for radar-aided UAV–UGV cooperative systems. *IEEE Access* 10, 50947–50958. doi: 10.1109/ACCESS.2022.3173806

Yang, L., Ding, B., Liao, W., and Li, Y. (2022). Identification of preisach model parameters based on an improved particle swarm optimization method for piezoelectric actuators in micro-manufacturing stages. *Micromachines* 13, 698. doi: 10.3390/mi13050698

Yang, X., Lei, K., Hu, L., Cao, X., and Huang, X. (2017). Eigenvalue ratio based blind spectrum sensing algorithm for multiband cognitive radios with relatively small samples. *Electron. Lett.* 53, 1150–1152. doi: 10.1049/el.2017.1658

Ye, S., Zhou, K., Zhang, C., Mohd Zain, A., and Ou, Y. (2022). An improved multi-objective cuckoo search approach by exploring the balance between development and exploration. *Electronics* 11, 704. doi: 10.3390/electronics11050704

Zeng, S., Tang, M., Sun, Q., and Lei, L. (2022). Robustness of interval-valued intuitionistic fuzzy reasoning quintuple implication method. *IEEE Access* 10, 8328–8338. doi: 10.1109/ACCESS.2022.3142766

Zhang, C., Zhou, K., Ye, S., and Zain, A. M. (2021). An improved cuckoo search algorithm utilizing nonlinear inertia weight and differential evolution for function optimization problem. *IEEE Access* 9, 161352–161373. doi: 10.1109/ACCESS.2021.3130640

Zhang, H., Li, P., Jin, H., Bi, R., and Xu, D. (2022a). Nonlinear wave energy dissipator with wave attenuation and energy harvesting at low frequencies. *Ocean Eng.* 266, 112935. doi: 10.1016/j.oceaneng.2022.112935

Zhang, X., Zhou, K., Li, P., Xiang, Y., Zain, A. M., and Sarkheyli-Hägele, A. (2022b). An improved chaos sparrow search optimization algorithm using adaptive weight modification and hybrid strategies. *IEEE Access* 10, 96159–96179. doi: 10.1109/ACCESS.2022.3204798

Zhang, Y. (2022). Tri-projection neural network for redundant manipulators. *IEEE Trans. Circ. Syst. II Express Briefs* 69, 4879–4883. doi: 10.1109/TCSII.2022.3189664

Zhang, Y., Chen, D., Guo, D., Liao, B., and Wang, Y. (2015). On exponential convergence of nonlinear gradient dynamics system with application to square root finding. *Nonlinear Dyn.* 79, 983–1003. doi: 10.1007/s11071-014-1716-3

Zhang, Y., Li, S., Kadry, S., and Liao, B. (2018a). Recurrent neural network for kinematic control of redundant manipulators with periodic input disturbance and physical constraints. *IEEE Trans. Cybernet.* 49, 4194–4205. doi: 10.1109/TCYB.2018.2859751

Zhang, Y., Li, S., Weng, J., and Liao, B. (2022c). Gnn model for time-varying matrix inversion with robust finite-time convergence. *IEEE Trans. Neural Netw. Learn. Syst.* doi: 10.1109/TNNLS.2022.3175899

Zhang, Y., Li, W., Liao, B., Guo, D., and Peng, C. (2014). Analysis and verification of repetitive motion planning and feedback control for omnidirectional mobile manipulator robotic systems. *J. Intell. Robot. Syst.* 75, 393–411. doi: 10.1007/s10846-014-0022-0

Zhang, Y., Qiu, B., Liao, B., and Yang, Z. (2017). Control of pendulum tracking (including swinging up) of IPC system using zeroing-gradient method. *Nonlinear Dyn.* 89, 1–25. doi: 10.1007/s11071-017-3432-2

Zhang, Y., and Yang, X. (2020). A 36 gb/s wireline receiver with adaptive CTLE and 1-tap speculative dfe in 0.13 μm bicmos technology. *IEICE Electron. Express* 17, 20200009. doi: 10.1587/elex.17.20200009

Zhang, Y., Zhang, J., and Weng, J. (2022d). Dynamic moore-penrose inversion with unknown derivatives: gradient neural network approach. *IEEE Trans. Neural Netw. Learn. Syst.* doi: 10.1109/TNNLS.2022.3171715

Zhang, Z., Deng, X., Qu, X., Liao, B., Kong, L.-D., and Li, L. (2018b). A varying-gain recurrent neural network and its application to solving online time-varying matrix equation. *IEEE Access* 6, 77940–77952. doi: 10.1109/ACCESS.2018.2884497

Zhang, Z., Fu, T., Yan, Z., Jin, L., Xiao, L., Sun, Y., et al. (2018c). A varying-parameter convergent-differential neural network for solving joint-angular-drift problems of redundant robot manipulators. *IEEE/ASME Trans. Mechatron.* 23, 679–689. doi: 10.1109/TMECH.2018.2799724

Zhang, Z., Kong, L., Zheng, L., Zhang, P., Qu, X., Liao, B., et al. (2020). Robustness analysis of a power-type varying-parameter recurrent neural network for solving time-varying QM and QP problems and applications. *IEEE Trans. Syst. Man Cybernet. Syst.* 50, 5106–5118. doi: 10.1109/TSMC.2018.2866843

Zhang, Z., Zheng, L., Li, L., Deng, X., Xiao, L., and Huang, G. (2018d). A new finite-time varying-parameter convergent-differential neural-network for solving nonlinear and nonconvex optimization problems. *Neurocomputing* 319, 74–83. doi: 10.1016/j.neucom.2018.07.005

Zhang, Z., Zheng, L., and Wang, M. (2019). An exponential-enhanced-type varying-parameter RNN for solving time-varying matrix inversion. *Neurocomputing* 338, 126–138. doi: 10.1016/j.neucom.2019.01.058

Zhao, H., Zhang, H., Bi, R., Xi, R., Xu, D., Shi, Q., et al. (2020). Enhancing efficiency of a point absorber bistable wave energy converter under low wave excitations. *Energy* 212, 118671. doi: 10.1016/j.energy.2020.118671

Zhou, K., Gui, W., Mo, L., and Zain, A. M. (2018a). A bidirectional diagnosis algorithm of fuzzy petri net using inner-reasoning-path. *Symmetry* 10, 192. doi: 10.3390/sym10060192

Zhou, K., Mo, L., Ding, L., and Gui, W.-H. (2018b). An automatic algorithm to generate a reachability tree for large-scale fuzzy petri net by and/or graph. *Symmetry* 10, 454. doi: 10.3390/sym10100454

Zhou, K., Mo, L., Jin, J., and Zain, A. M. (2019). An equivalent generating algorithm to model fuzzy petri net for knowledge-based system. *J. Intell. Manufact.* 30, 1831–1842. doi: 10.1007/s10845-017-1355-x

Zhou, K., Zain, A. M., and Mo, L.-P. (2015). A decomposition algorithm of fuzzy petri net using an index function and incidence matrix. *Expert Syst. Appl.* 42, 3980–3990. doi: 10.1016/j.eswa.2014.12.048

Zhou, P., Tan, M., Ji, J., and Jin, J. (2022). Design and analysis of anti-noise parameter-variable zeroing neural network for dynamic complex matrix inversion and manipulator trajectory tracking. *Electronics* 11, 824. doi: 10.3390/electronics11050824

Zhuo, D., and Cao, H. (2021). Fast sound source localization based on SRP-PHAT using density peaks clustering. *Appl. Sci.* 11, 445. doi: 10.3390/app11010445

Zhuo, D., and Cao, H. (2022). Damage identification of bolt connection in steel truss structures by using sound signals. *Struct. Health Monitor.* 21, 501–517. doi: 10.1177/14759217211004823

frontiers | Frontiers in Neurorobotics

TYPE Original Research
PUBLISHED 05 May 2023
DOI 10.3389/fnbot.2023.1178006

An advanced bionic knee joint mechanism with neural network controller

Changxian Xu, Zhongbo Sun*, Chen Wang, Xiujun Wu, Binglin Li and Liming Zhao*

Department of Control Engineering, Changchun University of Technology, Changchun, China

OPEN ACCESS

EDITED BY
Long Jin,
Lanzhou University, China

REVIEWED BY
Jiliang Zhang,
The University of Sheffield, United Kingdom
Bo Zhao,
Beijing Normal University, China

*CORRESPONDENCE
Liming Zhao
✉ zhaoliming904035@163.com
Zhongbo Sun
✉ zhongbosun2012@163.com

RECEIVED 02 March 2023
ACCEPTED 27 March 2023
PUBLISHED 05 May 2023

CITATION
Xu C, Sun Z, Wang C, Wu X, Li B and Zhao L (2023) An advanced bionic knee joint mechanism with neural network controller. *Front. Neurorobot.* 17:1178006.
doi: 10.3389/fnbot.2023.1178006

In this article, a tensegrity-based knee mechanism is studied for developing a high-efficiency rehabilitation knee exoskeleton. Moreover, the kinematics and dynamics models of the knee mechanism are explored for bringing about further improvement in controller design. In addition, to estimate the performance of the bionic knee joint, based on the limit function of knee patella, the limit position functionality of the bionic knee joint is developed for enhancing the bionic property. Furthermore, to eliminate the noise item and other disturbances that are constantly generated in the rehabilitation process, a noise-tolerant zeroing neural network (NTZNN) algorithm is utilized to establish the controller. This indicates that the controller shows an anti-noise performance; hence, it is quite unique from other bionic knee mechanism controllers. Eventually, the anti-noise performance and the calculation of the precision of the NTZNN controller are verified through several simulation and contrast results.

KEYWORDS

tensegrity, bionic knee joint, kinematics, dynamics, noise-tolerant zeroing neural network model

1. Introduction

Rigid–flexible coupling robot technology has broad application prospects in medical diagnosis, pipeline fault detection, bionic structure manufacturing, and other fields. The tensegrity structure is an important part of this technology because of its lightweight and deployable characteristics. In the process of rehabilitation training, due to the symptoms of hemiplegia caused by stroke or cerebral hemorrhage in the patient, the rehabilitation training of the human knee joint becomes quite important. The knee joint can be regarded as a strongly coupled structure that is composed of bones, muscles, and ligaments. Hence, the components of a knee joint cannot be simply mapped to the traditional rigid linkage structure. More importantly, the motion characteristics of the knee joint should be analyzed when the movement takes place (Oshkour et al., 2011). Therefore, a bionic knee joint structure based on the principle of bionics can be constructed using the rigid–flexible coupling tensegrity structure.

The lower limb rehabilitation training of several rehabilitation robots has been analyzed in Arsenault and Gosselin (2005, 2006a,b, 2009); Vasquez and Correa (2007); Murray et al. (2015); Esquenazi and Talaty (2019); Nicholson-Smith et al. (2020), and Muralidharan and Wenger (2021). Yet, these robots have not been analyzed from the perspective of bionics. Since the tensegrity structure is considered to be a rigid–flexible coupling mechanism in Jung et al. (2018) and Liu et al. (2020), the problem has been considered from the viewpoint of bionics mechanism, but the dynamics analysis has not been carried out due to structural complexity. In Collins et al. (2015), Sankai and Sakurai (2018), Fitzsimons et al. (2019), and Kim et al. (2020), the wearable exoskeletons, which can be utilized for the patient rehabilitation process with upper and lower limb disabilities, have been established. Two bionic robots based on the ankle joint and the knee joint have been studied in Sun et al. (2019) and Zhang et al. (2020). These two bionic robots have been formatted as the ankle and knee joint tensegrity structures based on the human body constitution. However, owing to structural complexity, the dynamics models are not studied on a temporary basis. Therefore, when faced with a complex environment, these two bionic tensegrity structures may not meet the practical requirement. For the purpose of implementing the actual rehabilitation training scenario, the interference caused by external environments and patients, such as the mechanical manufacturing errors and the static friction between the rehabilitation robot with patients, cannot be avoided. As a result, the bionic tensegrity structure based on the dynamics analysis of human lower limb joints under noise environment is of great significance for further research of bionic human joints.

Considering the fact that during the human lower limb rehabilitation process, the torque, which is produced by the knee joint, cannot be ignored. In the different rehabilitation processes, the knee joint produces different knee torques. These knee torques should be considered in the design of dynamics models, which can demonstrate the influence of human knee forces on the bionic knee mechanism during the movement (Rifai et al., 2013, 2016). In addition, noise is unavoidable in the process of a bionic knee joint movement. In the field of anti-noise algorithm, the NTZNN algorithm has shown its advantages in the parallelly distributed computing and anti-noise fields (Hehne, 1990; Jin et al., 2017, 2018; Sun et al., 2020; Shi et al., 2021; Wei et al., 2021). In this article, the error caused by the actual trajectory and the desired trajectory can be seen as a non-linear objective function. Furthermore, the kinematics and dynamics of the tensegrity mechanism are studied. In addition, the limited function of the knee is realized by the mechanical design, for the purpose of showing the bionic performance of a knee joint tensegrity structure. The article is formulated as follows. In Section 2, it describes the structure of the human knee joint and the establishment process of the bionic knee joint tensegrity structure mapping model. The kinematics of the proposed structural mechanism are presented in Section 3. The dynamics model and the description of the NTZNN controller are proposed in Section 4. Simulation results in Section 5 prove that the bionic knee joint tensegrity structure is effective under the noise condition. Finally, in Section 6, the conclusion and future study are discussed. At the end of this paragraph, the main contributions of the article are summarized as follows.

1. A bionic knee joint tensegrity mechanism is proposed and studied. Furthermore, the limit position functionality of the knee joint is achieved through a mechanical design. In addition, the NTZNN model has shown its efficiency in designing a controller with the distractions of noise items.
2. A series of simulation and contrast results with the proportional integral differential (PID) controller are presented to prove the accuracy, computational efficiency, and the anti-noise performance of the NTZNN controller.

2. Establishment process of a bionic knee joint

In this section, by analyzing the muscles, bones, and ligaments of a knee joint, the tissues of a knee joint are simplified into one component that has the same function during the movement. In addition, a bionic knee joint structure based on the tensegrity structure is established according to the characteristics of a human lower limb. Based on the principle of bionics, the physical characteristics of the bionic knee joint, such as the limit self-locking function and muscle elasticity coefficient, are considered in the design process of a bionic knee joint.

2.1. Structural description of the knee joint

To establish the bionic knee joint mapping model, there is demand to investigate the structure of the human knee joint in detail. Therefore, in this subsection, the human knee joint is analyzed for further research. It is crucial to notice that only sagittal motions are considered in this article. Hence, the use of a human knee joint is mainly employed in the sagittal plane of the lower limb movement, such as going up- and downstairs, squatting, and jumping.

The knee includes four bones, the lower part of the femur, the upper part of the tibia, the upper part of the fibula, and the patella. Femur, tibia, and fibula act as weight bearing bones and force transfer during the lower limb movement. In addition, the patella plays a limiting role in preventing the lower limb from overextending during movement, thus avoiding injury to the human body. Therefore, the patella location-restricted self-locking function is the key function of bionic joint tensegrity. Furthermore, due to the knee bearing the responsibility of supporting the body weight, its stiffness is higher. Thus, the skeleton of the knee joint can be regarded as the strut of a tensegrity structure, which indicates that the stiffness of a strut is infinite compared with the cable. The muscles and ligaments of the human lower limb are responsible for generating and transferring the load. Knee muscles can be divided into four groups according to their role in the lower limb movement. More importantly, the deformation of a muscle relative to the external load is shown in Figure 1. The muscle viscoelastic coefficient is similar to the spring damping coefficient, which should be considered in the stage of elastic range. To a certain degree, the bionics performance of the knee tensegrity mechanism can be realized by considering the viscoelastic coefficient.

2.2. Establishment process of the human knee joint mapping model

The knee joint mapping model and the bionic tensegrity structure are constructed in this subsection. For reducing the human tissue structure into a low-degree-of-freedom tensegrity structure, the strategy is to simplify the knee with basically the same function into the one structure. Under this strategy, the hamstring and tibialis anterior muscles are reduced to one muscle. The sartorius, semimembranosus, gracilis, and semitendinosus can

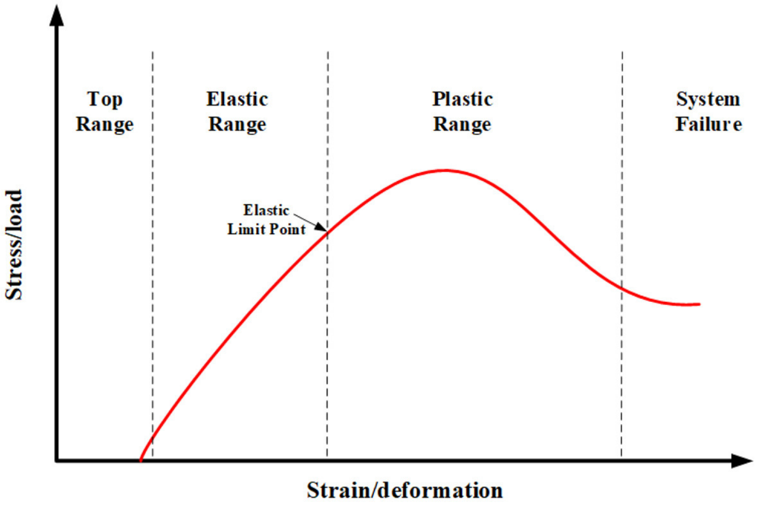


FIGURE 1
The relationship between the knee joint deformation with an external load (Bahr and Maehlum, 2003).

be seen as one muscle. Furthermore, the quadriceps is simplified to a muscle. In addition, the gastrocnemius could be thought of as a muscle. As regards the bone and the bone-like tissue, the fibula and tibia are decreased to a single bone for the reason of their similar functionality. Due to their peculiar function, the patella and tissues that perform the same function are simplified into two struts. The bionic patellar groove, which can also be called as the pulley groove, is established to implement the ultimate self-locking function of the knee joint. A limiting device is constructed on the bionic patella groove. It prevents the pulley from going off course as it slides through the bionic patellar groove. The self-locking function of the bionic knee joint is realized through the aforementioned mechanism design ultimately. As regards the bionic knee joint structure, the rotating pair and the first strut can be seen as the simplified bionic patella structures. For the sake of simplifying the complexity of a bionic patella mechanism, we should also ensure that the bionic mechanism should realize the bionic purpose. The rotation pair should be fixed for limiting the bionic knee extension movement under the action of external forces.

3. Analyses of the bionic knee joint

3.1. Description of a bionic knee joint mechanism

From the viewpoint of bionics, the patella and similar functional tissues perform two main biological functions during locomotion. In the first place, it distributes the pressure more widely over the femur, with the strategy of increasing the contact range between the patellar tendon and the patella. In the second place, it helps in knee extension by creating a forward displacement of the quadriceps tendon throughout the range of motion. However, the range of motion of the patella is fairly small, relative to the overall motion of the knee joint from full flexion to full extension (Hehne, 1990). Furthermore, when the knee joint moves, the displacement of the patella leading edge is not obvious when compared with the femur, fibula, and tibia. As a result, its primary function is to protect the quadriceps femoral tendon. For the sake of simplifying the complexity of the bionic knee structure, which also reduces the degree of structural freedom, as a result, it is convenient to analyze the dynamics model of the bionic knee joint structure in the next step. Moreover, for achieving the self-locking function of the patella and for the purpose of preventing knee hyperextension, the revolute joint pair is fixed to the bionic patella. The pulley is slid in the bionic patella groove, which is aimed to finish the self-locking function of the knee joint during the lower limb movement. The range of knee flexion angles for a healthy adult is approximately 130^o to 140^o, but the stroke patients cannot complete the entire motion range. However, the range of motion of the affected limb is increased when the affected limb's physical condition is improved during the rehabilitation process. Therefore, the length of a bionic patellar groove can be changed to satisfy the different rehabilitation training stages.

3.2. Singular configuration

Singular configuration refers to the case where degeneracy occurs between the input and output variables of the structure (Arsenault and Gosselin, 2005). However, due to the limiting properties provided by the strut *CF*, *DE*, and bionic patellar groove, the bionic knee tensegrity structure may stop moving before reaching the singular configuration. Thus, the singular configuration is reached when the knee extension is the upper working boundary of the tensegrity structure. However, when the movement of the mechanism takes place, this situation should be avoided. In this case, the tensegrity system is degenerated, which may cause the tensegrity system to collapse. Furthermore, the situation is similar to the undue knee joint movements that could happen in real life rehabilitation.

3.3. Working curve

In the design process of a tensegrity structure, it is very important to study the working curve of the mechanism. In this subsection, the working curves of nodes C and D are obtained through the ADAMS software kinematic simulation, which can get the working spaces of angles θ and γ. Since two sets of linkage mechanisms are axially symmetric about the y axis, the operating curves of the two nodes are identical. In the kinematic simulation, the external forces are perpendicular to the x axis, which are acting on the nodes C and D; hence, the operating curve goes from the initial self-equilibrium state to the limit position when θ is equal to 90^o. It can be seen from the kinematic simulation that the limiting mechanism based on the principle of bionics can prevent the overextension of the bionic knee joint structure under the action of external forces on nodes C and D. However, the displacements of points A and B cannot be restricted through the two-link mechanism alone. It reflects the significance for the bionic patellar groove's constraint functionality when facing the movement at points A and B. Furthermore, the y_C and y_D decrease when the force direction is opposite to the previous situation, the circumstance corresponds to the knee flexion. The y_C and y_D will decrease to zero eventually, yet the situation should be avoided in the actual operating circumstance.

4. Dynamics model and controller

To exploit the efficiency of the bionic knee joint tensegrity structure in the rehabilitation process, the dynamics model and the NTZNN controller of the presented tensegrity bionic knee joint are developed and studied in this section.

4.1. Dynamics model

4.1.1. Hypotheses

The following hypotheses are proposed to derive the dynamics model of the tensegrity structure:

1. The gravitational potential energy is neglected for the purpose of reducing the dynamics model's complexity.
2. The springs are massless.
3. Each strut is a thin rod of w mass and the moment of inertia is $\frac{1}{12}wB^2$.
4. The spring is linearly damped with coefficients a_1, a_2, a_3, and a_4, in which a_1 is equal to a_2.

4.1.2. Equation form of the Lagrangian approach

As regards the tensegrity structure, it has two degrees of freedom, therefore, the β and γ are selected as the generalized coordinates. The dynamics model is developed by utilizing the Lagrangian approach, which is defined by

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial K}{\partial \dot{\mathbf{q}}} - \frac{\partial K}{\partial \mathbf{q}} + \frac{\partial P}{\partial \mathbf{q}} = \mathbf{f}, \tag{1}$$

where P and K express the potential and kinetic energies of the tensegrity structure, $\mathbf{f} = \left[f_1, f_2\right]^{\mathrm{T}}$ is the non-conservative force, and the $\mathbf{q}$ is equal to $[\beta, \gamma]^{\mathrm{T}}$. To reflect the influence on the dynamics model of viscoelasticity caused by muscle deformation, the non-conservative forces are formed in this subsection. The non-conservative forces correspond to the damping forces in the springs. In the tensegrity mechanism, the kinetic energy of the system is generated by the movement of the strut alone, thus the kinetic energy can be formatted as

$$K = w_1 B_2^2 \dot{\beta}^2 + \frac{1}{3} w_1 B_1^2 (\dot{\beta} + \dot{\gamma})^2 + w_1 B_1 B_2 cos\gamma \dot{\beta}(\dot{\beta} + \dot{\gamma}) + \frac{1}{3} w_2 B_2^2 \dot{\beta}^2, \tag{2}$$

where w_1 and w_2 are the masses of struts B_1 and B_2, individually.

In addition, the potential energy could be defined as:

$$P = k_1\left(\sqrt{(B_1\cos\alpha - 2(B_3 + B_2\cos\beta))^2 + (B_2\sin\beta)^2} - z_{01}\right)^2 + \frac{1}{2}k_3(2(B_3 + B_2\cos\beta) - z_{03})^2 + \frac{1}{2}k_4(2(B_1\cos\alpha - (B_3 + B_2\cos\beta)) - z_{04})^2, \tag{3}$$

where the subentry potential energy P_1 is equal to P_2. The z_{01}, z_{03}, and z_{04} are the initial lengths of the springs. The non-conservative force caused by spring damping is expressed as

$$\begin{aligned} f_1 &= -2c_1\dot{z}_1\frac{\partial z_1}{\partial \beta} - c_3\dot{z}_3\frac{\partial z_3}{\partial \beta} - c_4\dot{z}_4\frac{\partial z_4}{\partial \beta} \\ f_2 &= -2c_1\dot{z}_1\frac{\partial z_1}{\partial \gamma} - c_3\dot{z}_3\frac{\partial z_3}{\partial \gamma} - c_4\dot{z}_4\frac{\partial z_4}{\partial \gamma}, \end{aligned} \tag{4}$$

where the z_1, z_3, and z_4 are the presented lengths of the springs. As shown in Figure 1, the coefficient of muscle elasticity in the elastic range is similar to the coefficient of spring damping. The muscles are similar to the springs in the bionic knee joint. When muscles are deformed, the resistance produced by the friction between muscle fibers sticks to the extension and contraction of muscles. Consequently, the bionic performance of the bionic knee joint structure can be achieved by considering the elastic damping in the dynamics modeling process.

Hence, the dynamics model can be formatted as follows:

$$\mathbf{M}\ddot{\mathbf{q}} + \mathbf{H}\dot{\mathbf{q}}_{qh} + \mathbf{G}\dot{\mathbf{q}}_{qg} + \mathbf{C}\dot{\mathbf{q}} + \mathbf{T} + \mathbf{u} + \boldsymbol{\tau}_R = 0, \tag{5}$$

where $\dot{\mathbf{q}}_{qh}=\left[\dot{\beta}^2, \dot{\gamma}^2\right]^{\mathrm{T}}$, $\dot{\mathbf{q}}_{qg}=\left[\dot{\beta}\dot{\gamma}, \dot{\gamma}\dot{\beta}\right]^{\mathrm{T}}$, $\boldsymbol{\tau}_R$, and $\mathbf{u}$ are the knee torque and control law, the matrix $\mathbf{C}$ has relations with the non-conservative force, and $\mathbf{T} = [T_1, T_2]^{\mathrm{T}}$ is the matrix that is associated with the potential energy.

Detailedly,

$$W_{1,1} = 2w_1B_1^2 + \frac{2}{3}w_1B_2^2 + \frac{2}{3}w_2B_2^2 + 2w_1B_1B_2\cos\gamma, \tag{6}$$

$$W_{1,2} = W_{2,1} = \frac{2}{3}w_1B_1^2 + w_1B_1B_2\cos\gamma, \tag{7}$$

$$W_{2,2} = \frac{2}{3}w_1B_1^2, \tag{8}$$

moreover,

$$\mathbf{H} = \begin{bmatrix} 0 & -w_1B_1B_2\sin\gamma \\ w_1B_1B_2\sin\gamma & 0 \end{bmatrix}, \tag{9}$$

and,

$$\mathbf{G} = \begin{bmatrix} -2w_1 B_1 B_2 \sin\gamma & 0 \\ 0 & 0 \end{bmatrix}. \tag{10}$$

4.2. NTZNN controller

A continuous-time NTZNN model is utilized to design the control law in this subsection. In the process of the operation of the mechanism, the noises, which may include mechanical structure error, mechanical vibration, friction between components, feedback signal noise, external static friction and other factors, are inevitable items. In addition, the knee torque, which is generated by the human knee during rehabilitation, should be considered in the dynamics modeling process. In the human lower limb recovery process, different lower limb rehabilitation stages may cause different torques which are produced by the knee. For example, in the early stage of rehabilitation training, lower limb hemiplegia that is caused by stroke and other diseases may lead to an uncoordinated movement of lower limbs, which make lower limbs unable to move according to the patient's real intention. The actual lower limb movement trajectory may be in conflict with the rehabilitation robot. In addition, there is a special rehabilitation stage, which corresponds to be deprived of the nerve conduction function between the patient's central nervous system and the lower limb skeletal muscles. It could also be considered as the passive rehabilitation stage of a patient who has received a lower limb joint surgery or a total knee replacement, and in these circumstances, the knee torque $\boldsymbol{\tau}_R$ is very small when compared with other situations (Cao and Huang, 2020). Therefore, in the modeling process, the knee torque could not be overlooked, and the knee torque $\boldsymbol{\tau}_R$ should be considered in the modeling process. The knee torque $\boldsymbol{\tau}_R$ could be seen as a constant torque, for the reason that in the same rehabilitation stage, the knee torque is roughly the same.

In this subsection, the problem is formatted as:

$$\boldsymbol{\phi}(y(p)) = 0 \in \mathbb{R}, p \in [0, +\infty), \tag{11}$$

furthermore,

$$\frac{\mathrm{d}\boldsymbol{\phi}(y(p))}{\mathrm{d}p} = \frac{\partial \boldsymbol{\phi}(y(p))}{\partial t} + \frac{\partial \boldsymbol{\phi}(y(p))}{\partial y(p)}\frac{\mathrm{d}y(p)}{dp} = \dot{\boldsymbol{\phi}}_p(y(p)) + \mathbf{R}(y(p))\frac{\mathrm{d}y(p)}{\mathrm{d}p}, \tag{12}$$

where $\mathbf{R}(y(p))$ is equal to $\partial\boldsymbol{\phi}(y(p))\big/\partial y(p)$.

An error function can be generalized as:

$$\mathbf{e}(p) = 0 - \boldsymbol{\phi}(y(p)). \tag{13}$$

Hence, a noise-suppressing zeroing dynamics model is defined by:

$$\dot{\mathbf{e}}(p) = -\beta\mathbf{e}(p) - \lambda\int_0^t \mathbf{e}(\delta)\mathrm{d}\delta, \tag{14}$$

where β and λ are positive constants. δ is the time interval. Eventually, a continuous-time NTZNN model which is polluted by noise is given as:

$$\dot{y}(p) = -\mathbf{R}^{-1}(y(p))(\beta\boldsymbol{\phi}(y(p)) + \dot{\boldsymbol{\phi}}_p(y(p)) + \lambda\int_0^t \boldsymbol{\phi}(y(\delta))\mathrm{d}\delta + \boldsymbol{\varepsilon}(p)), \tag{15}$$

which $\boldsymbol{\varepsilon}(p)$ is the noise item. In this subsection, considering the influence of noises and knee joint torques on the bionic knee joint control algorithm, an anti-noise ZNN model is established as the control algorithm to control the bionic knee joint dynamics model. To further study the NTZNN model, the theories are presented as follows.

Theorem 1. The $\phi(p)$ can be seen as a vector, which is to say that the time-varying vector $\Xi(p)$ can be managed through utilization of the NTZNN model global convergence from selecting the initial states ($\Xi_0 \neq 0 \in \mathbb{R}$) to the theoretical solution $\hat{\Xi}(p)$ randomly with constant noise ($R(p) = \tilde{R} \in \mathbb{R}$).

Proof The noise-polluted NTZNN model could be transformed based on the Laplace transformation, which can be formatted as follows

$$jy(j) - y(0) = -\Lambda y(j) - \frac{\iota}{j}y(j) + R(j).$$

As a result, the equation could be formed as

$$y(j) = \frac{j[y(0) + R(j)]}{j^2 + j\Lambda + \iota}. \tag{16}$$

Furthermore, the transfer function of equation (16) should be formatted as $j/(j^2 + j\Lambda + \iota)$. In addition, the $j_1 = (-\Lambda + \sqrt{\Lambda^2 - 4\iota})/2$ and $j_2 = (-\Lambda - \sqrt{\Lambda^2 - 4\iota})/2$ are poles of the transfer function. Moreover, on account of $\Lambda > 0$ and $\iota > 0$, the poles of the transfer function lie in the left half-plane, which can testify that the time-varying problem, which is polluted with the constant noise $R(p)$, is stable. In addition, for the reason of the noise item is constant, hence, $R(j) = \tilde{R}/j$. In summary, the following result can be defined as

$$\lim_{p\to\infty} y(p) = \lim_{s\to 0} jy(j) = \lim_{j\to 0}\frac{j^2[y(0) + \frac{\tilde{R}}{j}]}{j^2 + j\Lambda + \iota} = 0.$$

The proof is thus complete.

Theorem 2. When $\phi(p)$ can be seen as a vector, it is to say that the time-varying vector $\Xi(p)$ can be managed through utilization of the NTZNN model global convergence from selecting the initial states ($\Xi_0 \neq 0 \in \mathbb{R}$) to the solution $\hat{\Xi}(p)$ with linear noise ($R(p) = p\tilde{R} \in \mathbb{R}$).

Proof For the reason of the Laplace transformation, the NTZNN model with linear noise polluted ($R(p) = t\tilde{R}$) should be defined as

$$jy(j) = y(0) - \Lambda y(j) - \frac{\iota}{j}y(j) + \frac{\tilde{R}}{j^2}, \tag{17}$$

where $\tilde{R}/j^2$ is the Laplace transformation of $R(p)$. Hence, the following results could be formatted through investigation of the final value theorem

$$\lim_{t\to\infty} y(p) = \lim_{j\to 0}\frac{j^2[y(0) + \frac{\tilde{R}}{j^2}]}{j^2 + j\Lambda + \iota} = \frac{\tilde{R}}{\iota}.$$

As a result, a conclusion that $\lim_{p\to\infty} y(p) \to 0$ with $\iota \to \infty$ could be drawn. The proof is complete.

5. Experiments and analysis

In this section, through the experiments, the effectiveness of the bionic knee joint is verified under the interference of noise items.

5.1. The performance of a bionic knee joint tensegrity structure in different stages of rehabilitation

The moment and fixed noise of the knee joint in different rehabilitation stages are considered in the experiment to prove the anti-noise performance of the NTZNN controller and the accuracy of the bionic knee joint dynamics model. In the said experiments, three kinds of knee torques are proposed to represent the forces that are generated in different recovery stages. The three stages, namely, resistance rehabilitation stage, auxiliary rehabilitation stage, and passive rehabilitation stage, are distinguished by the knee joint torques, which are -40, 150, and 0 $\mathrm{N}\cdot\mathrm{m}$, respectively (Zhao and Xu, 2011). Furthermore, to reflect the superiority of the NTZNN algorithm in an anti-noise field, a kind of mixed noise, which is formed by constant noise, linear noise, and random noise is presented in this subsection. The fixed noise is defined by

$$\boldsymbol{\varepsilon}(t) = \boldsymbol{\eta} + \boldsymbol{\kappa}t + \boldsymbol{\mu}(t), \tag{18}$$

in which $\boldsymbol{\eta}$ is the constant noise, $\boldsymbol{\kappa}t$ is the linear noise, and $\boldsymbol{\mu}(t)$ is the random noise. $\boldsymbol{\eta}$, $\boldsymbol{\kappa}$, and $\boldsymbol{\mu}$ are the coefficients.

The desired trajectory of the bionic knee joint in the experiment is acquired and fitted by ADAMS software. Therefore, the desired trajectory can delegate the real motion trajectory of the bionic knee joint in rehabilitation training. The motion trajectory can be defined as follows

$$\theta_d = 1.33 - 0.2186 \times \cos(0.03749t) - 0.007953 \times \sin(0.03749t), \tag{19}$$

$$\begin{aligned}\gamma_d = 3.159 - 0.4651 \times \cos(0.03748t) - 0.01676 \times \sin(0.03748t) \\ - 0.1284 \times \cos(0.07496t) - 0.008771 \times \sin(0.07496t).\end{aligned} \tag{20}$$

5.1.1. Resistance rehabilitation stage

It is assumed that the patient is in the resistance rehabilitation stage. Although the patient can move the affected limb, it cannot carry out a series of rehabilitation activities completely according to the patient's real movement intention. The actual movement trajectory of the affected limb may encounter human–machine confrontation with the rehabilitation robot due to the uncoordinated movement of the affected limb. In addition, a series of rehabilitation training actions cannot be repeated for a long time due to muscle atrophy of the affected limb and various other reasons. Therefore, in the resistance rehabilitation stage, the knee joint torque generated by the affected knee joint is defined as a negative value, where the knee torque $\boldsymbol{\tau}_R$ is equal to -40 $\mathrm{N}\cdot\mathrm{m}$. The desired trajectory is utilized to explore the performance of the NTZNN controller. Figure 2 shows the position error between the actual trajectory and the desired trajectory of the bionic knee joint angle when using the NTZNN controller. The actual trajectory can converge to the desired trajectory using the NTZNN controller rapidly. The interference of internal and external noises to the model is considered during the design process of the NTZNN controller. The experimental results show that the fixed noise can be suppressed using the NTZNN model, which proves that the NTZNN algorithm has strong robustness and noise suppression ability. Although at the initial stage, there is an oscillation between the expected trajectory and the actual trajectory, nevertheless, with the increase in iterations, the error between the desired trajectory and the actual trajectory decreases and could reach the level of 1×10^{-4} gradually.

5.1.2. Assistance rehabilitation stage

The physical condition of the affected limb will improve after a period of rehabilitation training. In this process, the affected limb of the patient moves smoothly, but the affected limb is generally unable to produce enough torque to carry out rehabilitation training in accordance with the requirements of rehabilitation training. Hence, patients still need the bionic knee to provide an additional torque to assist the affected limb to complete rehabilitation training in the assistance rehabilitation phase. The knee torque that a healthy adult can produce is around 170 $\mathrm{N}\cdot\mathrm{m}$ to 300 $\mathrm{N}\cdot\mathrm{m}$. Although the affected limb can produce more torques in the auxiliary rehabilitation stage, it is still smaller than the normal torque. Thus, the knee joint torque is set as 150 $\mathrm{N}\cdot\mathrm{m}$ in the assistance rehabilitation stage. As shown in Figure 3, the fixed noise and knee torque 150 $\mathrm{N}\cdot\mathrm{m}$ are taken into account in the designing process of the NTZNN controller. The experimental results show that the NTZNN model could suppress the noise available, which makes the controller to be provided with robustness and anti-noise performances. In the assistance rehabilitation stage, the main objective for the controller is to manage the bionic knee movement under noise pollution. In addition, the purposes for using a bionic knee are to enhance the muscle strength by assisting with rehabilitation exercises and to facilitate the reconstruction of the somatosensory stimuli according to rehabilitation goals. The experiments have proved that under the noise pollution, the NTZNN controller could suppress the fixed noise and control the bionic knee to assist the patient to complete the rehabilitation goals, which demonstrate the accuracy and effectiveness of the NTZNN approach.

5.1.3. Passive rehabilitation stage

To verify the versatility of the bionic knee joint, that is, rehabilitation training can be completed under various circumstances, this subsection designs a passive rehabilitation stage. This situation applies to patients who have undergone knee surgery or total knee replacement surgery. Therefore, the rehabilitation training action of the affected limb is completely driven by the bionic knee joint. Compared with the torque generated in other rehabilitation stages, the knee joint torque in the passive rehabilitation stage is very small, so the knee joint torque is approximately equal to zero. Since the knee joint torque is zero,

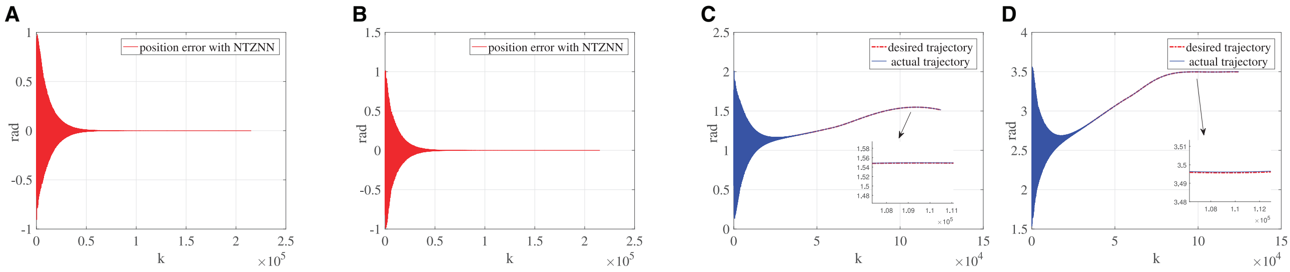


FIGURE 2
Under the circumstance of $\boldsymbol{\tau}_R$ is equal being −40, **(A)** is the error between the desired trajectory and the actual trajectory of θ, **(B)** the error between the desired trajectory and the actual trajectory of γ, **(C)** the desired the trajectory and the actual trajectory of θ and **(D)** the desired trajectory and the actual trajectory of γ.

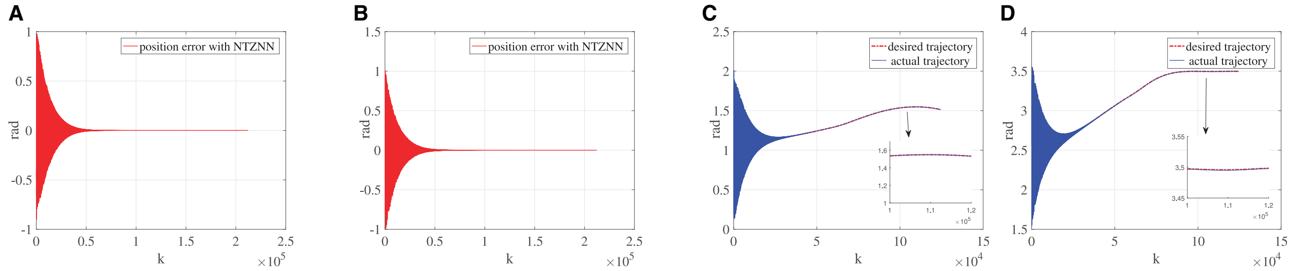


FIGURE 3
Under the circumstance of $\boldsymbol{\tau}_R$ being equal to 150, **(A)** is the error between the desired trajectory and the actual trajectory of θ, **(B)** the error between desired trajectory and the actual trajectory of γ, **(C)** the desired trajectory and the actual trajectory of θ and **(D)** the desired trajectory and the actual trajectory of γ.

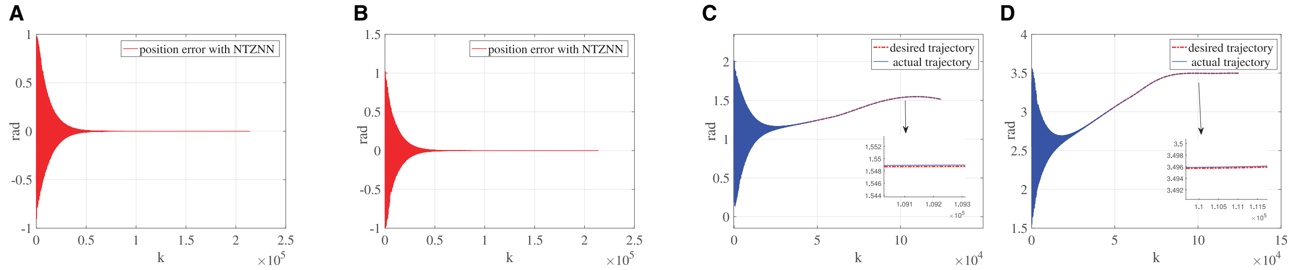


FIGURE 4
Under the circumstance of $\boldsymbol{\tau}_R$ being equal to 0, **(A)** is the error between the desired trajectory and the actual trajectory of θ, **(B)** the error between the desired trajectory and the actual trajectory of γ, **(C)** the desired trajectory and the actual trajectory of θ and **(D)** the desired trajectory and the actual trajectory of γ.

the only interference in the dynamics model is the fixed noise during passive rehabilitation. As shown in Figure 4, the dynamics model can achieve a low error and the convergence speed is faster than other rehabilitation stages under the control of the NTZNN algorithm. It also verifies that the knee torque $\boldsymbol{\tau}_R$ should be seen as a disturbance torque during the movement of the bionic knee joint, which demonstrates the importance of the NTZNN controller's anti-noise performance.

5.2. Contrast experiments

To demonstrate the superiority of the NTZNN algorithm in the field of noise suppression, the PID algorithm is used as the controller to control the bionic knee joint dynamics model in the contrast experiments under the noise condition. In the actual rehabilitation training process, not only will the external environment cause interference to the rehabilitation training process, but the patient's own health conditions will also cause certain interference to the rehabilitation training process, such as an involuntary spasm of the affected limb. The experimental results show that, the position errors of bionic knee joint angles θ and γ will increase with the introduction of fixed noise and knee torque gradually. It could be seen from Figure 5 that, with the growing number of iterations, the position error between the desired trajectory and the actual trajectory increases to the extent that it can affect the operation of the bionic knee joint. The interferences of external environment and

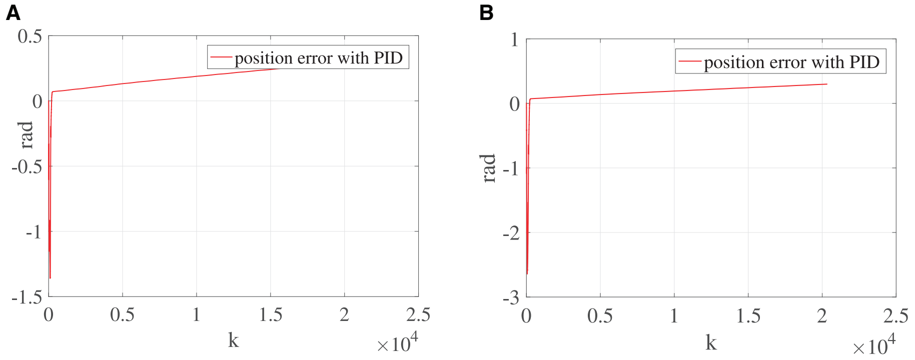


FIGURE 5
Under the circumstance of τ_R being equal to 150, **(A)** is the error between the desired trajectory and the actual trajectory of θ and **(B)** the error between the desired trajectory and the actual trajectory of γ.

patients to the rehabilitation training process are inevitable in the actual training situation under the interference of non-ideal factors. If the interferences to a rehabilitation training process are ignored when designing the control algorithm of a bionic knee joint, it may cause secondary injury to the affected limb during the rehabilitation process. Therefore, the NTZNN algorithm with an anti-noise ability offers great advantages in the design course of a bionic knee joint control algorithm. An NTZNN algorithm is established as the controller of a bionic knee joint dynamics model by analyzing the influence of the external noise and the knee joint torque on the bionic knee joint control algorithm in the actual rehabilitation training process. The experiments of three different rehabilitation stages and comparison experiments show that the NTZNN algorithm has significant advantages in suppressing non-ideal factors in rehabilitation training.

6. Conclusion

In this article, a bionic knee joint tensegrity structure in noise environment has been developed and studied from the viewpoint of principle of bionics. Moreover, the knee joint torques at different rehabilitation stages have been considered in the controller design process, so as to reflect the influence of the human knee force acting on the bionic knee joint tensegrity structure. The dynamics model of the bionic knee mechanism has been established by means of analyzing the kinetic energy and potential energy of the system. Eventually, the simulations and contrast results have shown that the NTZNN controller has advantages in noise suppression and computational efficiency. The main work in future would be to undertake research on the bionic hip joint structure with a remarkable bionic performance.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

CX and ZS: data curation and software. CX, ZS, and CW: conceptualization. CX, XW, BL, and LZ: methodology. XW, LZ, and CW: formal analysis. CX, CW, and XW: writing-original draft. ZS and LZ: supervision. ZS, CW, and LZ: validation. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported in parts by the National Natural Science Foundation of China under grants 61873304, 62173048, and 62106023, the Innovation and Entrepreneurship Talent Funding Project of Jilin Province 2022QN04, and the Changchun Science and Technology Project under grant 21ZY41.

Acknowledgments

The authors would like to thank the reviewers and the Technical Editor for their valuable comments and suggestions on revising this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Arsenault, M., and Gosselin, C. M. (2005). Kinematic, static, and dynamic analysis of a planar one-degree-of-freedom tensegrity mechanism. *Trans. ASME* 127, 1152–1160. doi: 10.1115/1.1913705

Arsenault, M., and Gosselin, C. M. (2006a). Kinematic, static and dynamic analysis of a planar 2-DOF tensegrity mechanism. *Mech. Mach. Theory* 41, 1072–1089. doi: 10.1016/j.mechmachtheory.2005.10.014

Arsenault, M., and Gosselin, C. M. (2006b). Kinematic, static, and dynamic analysis of a spatial three-degree-of-freedom tensegrity mechanism. *J. Mech. Design* 128, 1061–1069. doi: 10.1115/1.2218881

Arsenault, M., and Gosselin, C. M. (2009). Kinematic and static analysis of a 3-PUPS spatial tensegrity mechanism. *Mech. Mach. Theory* 44, 162–179. doi: 10.1016/j.mechmachtheory.2008.02.005

Bahr, R., and Maehlum, S. (2003). *Clinical Guide to Sports Injuries, Human Kinetics.*

Cao, Y., and Huang, J. (2020). Neural-network-based nonlinear model predictive tracking control of a pneumatic muscle actuator-driven exoskeleton. *IEEE/CAA J. Automat. Sin.* 7, 1478–1488. doi: 10.1109/JAS.2020.1003351

Collins, S. H., Wiggin, M. B., and Sawicki, G. S. (2015). Reducing the energy cost of human walking using an unpowered exoskeleton. *Nature* 522, 212–215. doi: 10.1038/nature14288

Esquenazi, A., and Talaty, M. (2019). Robotics for lower limb rehabilitation. *Phys. Med. Rehabil. Clin. N. Am.* 30, 385–397. doi: 10.1016/j.pmr.2018.12.012

Fitzsimons, K., Acosta, A. M., Dewald, J. P. A., and Murphey, T. D. (2019). Ergodicity reveals assistance and learning from physical human-robot interaction. *Sci. Robot.* 4, 60–79. doi: 10.1126/scirobotics.aav6079

Hehne, H. J. (1990). Biomechanics of the patellofemoral joint and its clinical relevance. *Clin. Orthop. Relat. Res.* 258, 73–85. doi: 10.1097/00003086-199009000-00011

Jin, L., Zhang, Y., Li, S., and Zhang, Y. (2017). Noise-tolerant ZNN models for solving time-varying zero-finding problems: a control-theoretic approach. *IEEE Trans. Automat. Control* 62, 992–997. doi: 10.1109/TAC.2016.2566880

Jin, L., Zhang, Y., and Qiu, B. (2018). Neural network-based discrete-time Z-type model of high accuracy in noisy environments for solving dynamic system of linear equations. *Neural Comput. Appl.* 29, 1217–1232. doi: 10.1007/s00521-016-2640-x

Jung, E., Ly, V., Cessna, N., Ngo, M. L., Castro, D., SunSpiral, V., et al. (2018). "Bio-inspired tensegrity flexural joints," in *2018 IEEE International Conference on Robotics and Automation* (Brisbane, QLD: IEEE), 5561–5566. doi: 10.1109/ICRA.2018.846102

Kim, K., Agogino, A. K., and Agogino, A. M. (2020). Rolling locomotion of cable-driven soft spherical tensegrity robots. *Soft Robot.* 7, 346–361. doi: 10.1089/soro.2019.0056

Liu, S., Li, Q., Wang, P., and Guo, F. (2020). Kinematic and static analysis of a novel tensegrity robot. *Mech. Mach. Theory* 149, 103788. doi: 10.1016/j.mechmachtheory.2020.103788

Muralidharan, V., and Wenger, P. (2021). Optimal design and comparative study of two antagonistically actuated tensegrity joints. *Mech. Mach. Theory* 159, 104249. doi: 10.1016/j.mechmachtheory.2021.104249

Murray, S. A., Ha, K. H., Hartigan, C., and Goldfarb, M. (2015). An assistive control approach for a lower-limb exoskeleton to facilitate recovery of walking following stroke. *IEEE Trans. Neural Syst. Rehabil. Eng.* 23, 441–449. doi: 10.1109/TNSRE.2014.2346193

Nicholson-Smith, C., Mehrabi, V., Atashzar, S. F., and Patel, R. V. (2020). A multi-functional lower- and upper-limb stroke rehabilitation robot. *IEEE Trans. Med. Robot. Bionics* 2, 549–552. doi: 10.1109/TMRB.2020.3034497

Oshkour, A., Osman, N. A., Davoodi, M., Bayat, M., Yau, Y., and Abas, W. W. (2011). "Knee joint stress analysis in standing," in *5th Kuala Lumpur International Conference on Biomedical Engineering*. p. 179–181. doi: 10.1007/978-3-642-21729-6_47

Rifai, H., Mohammed, S., Djouani, K., and Amirat, Y. (2016). Toward lower limbs functional rehabilitation through a knee-joint exoskeleton. *IEEE Trans. Control Syst. Technol.* 25, 1–8. doi: 10.1109/TCST.2016.2565385

Rifai, H., Mohammed, S., Hassani, W., and Amirat, Y. (2013). Nested saturation based control of an actuated knee joint orthosis. *Mechatronics* 23, 1141–1149. doi: 10.1016/j.mechatronics.2013.09.007

Sankai, Y., and Sakurai, T. (2018). Exoskeletal cyborg-type robot. *Sci. Robot.* 187, 1–9. doi: 10.1126/scirobotics.aat3912

Shi, T., Tian, Y., Sun, Z., Liu, K., Jin, L., and Yu, J. (2021). Noise-tolerant neural algorithm for online solving yang-baxter-type matrix equation in the presence of noises: a control-based method. *Neurocomputing* 424, 84–96. doi: 10.1016/j.neucom.2020.10.110

Sun, J., Song, G., Chu, J., and Ren, L. (2019). An adaptive bioinspired foot mechanism based on tensegrity structures. *Soft Robot.* 6, 778–789. doi: 10.1089/soro.2018.0168

Sun, Z., Shi, T., Wei, L., Sun, Y., Liu, K., and Jin, L. (2020). Noise-suppressing zeroing neural network for online solving time-varying nonlinear optimization problem: a control-based approach. *Neural Comput. Appl.* 32, 11505–11520. doi: 10.1007/s00521-019-04639-2

Vasquez, R. E., and Correa, J. C. (2007). "Kinematics, dynamics and control of a planar 3-DOF tensegrity robot manipulator," in *Proceedings of the ASME 2007 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. Volume 8: 31st Mechanisms and Robotics Conference, Parts A and B* (Las Vegas, NV: ASME), 855–866. doi: 10.1115/DETC2007-34975

Wei, L., Jin, L., Yang, C., Chen, K., and Li, W. (2021). New noise-tolerant neural algorithms for future dynamic nonlinear optimization with estimation on Hessian matrix inversion. *IEEE Trans. Syst. Man Cybernet. Syst.* 51, 2611–2623. doi: 10.1109/TSMC.2019.2916892

Zhang, W., Liu, L., and Song, G. (2020). Design of bionic knee joint mechanism based on tensegrity structure. *Eng. Struct.* 44, 98–104. doi: 10.16578/j.issn.1004.2539.2020.12.015

Zhao, H., and Xu, X. (2011). Torque parameters of human knee joint. *J. Clin. Rehabil. Tissue Eng. Res.* 15, 705–708. doi: 10.3969/j.issn.1673-8225.2011.04.033

frontiers | Frontiers in Neurorobotics

TYPE Original Research
PUBLISHED 14 June 2023
DOI 10.3389/fnbot.2023.1181864

OPEN ACCESS

EDITED BY
Long Jin,
Lanzhou University, China

REVIEWED BY
Fangwen Yu,
Tsinghua University, China
P. Chandra Shaker Reddy,
Geethanjali College of Institute of Engineering and Technology, India
Hao Xu,
Anhui University of Technology, China
Guadalupe Dorantes,
Autonomous University of San Luis Potosí, Mexico

*CORRESPONDENCE
Xue-Bo Jin
✉ jinxuebo@btbu.edu.cn

RECEIVED 08 March 2023
ACCEPTED 23 May 2023
PUBLISHED 14 June 2023

CITATION
Bai Y-T, Jia W, Jin X-B, Su T-L and Kong J-L (2023) Location estimation based on feature mode matching with deep network models. *Front. Neurorobot.* 17:1181864. doi: 10.3389/fnbot.2023.1181864

Location estimation based on feature mode matching with deep network models

Yu-Ting Bai[1,2], Wei Jia[1], Xue-Bo Jin[1,2]*, Ting-Li Su[1] and Jian-Lei Kong[1]

[1]School of Artificial Intelligence, Beijing Technology and Business University, Beijing, China, [2]Beijing Laboratory for Intelligent Environmental Protection, Beijing Technology and Business University, Beijing, China

Introduction: Global navigation satellite system (GNSS) signals can be lost in viaducts, urban canyons, and tunnel environments. It has been a significant challenge to achieve the accurate location of pedestrians during Global Positioning System (GPS) signal outages. This paper proposes a location estimation only with inertial measurements.

Methods: A method is designed based on deep network models with feature mode matching. First, a framework is designed to extract the features of inertial measurements and match them with deep networks. Second, feature extraction and classification methods are investigated to achieve mode partitioning and to lay the foundation for checking different deep networks. Third, typical deep network models are analyzed to match various features. The selected models can be trained for different modes of inertial measurements to obtain localization information. The experiments are performed with the inertial mileage dataset from Oxford University.

Results and discussion: The results demonstrate that the appropriate networks based on different feature modes have more accurate position estimation, which can improve the localization accuracy of pedestrians in GPS signal outages.

KEYWORDS

location estimation, feature extraction, mode classification, deep networks, location system

1. Introduction

In the information age, navigation technology is constantly innovated in national defense and the lives of people and society (Jin et al., 2023). Location estimation and positioning are based on sensors, communication, and electronic control technology to connect resources and information (Dong et al., 2021). Satellite navigation has been the mainstream of location estimation and positioning. However, the navigation signal will be lost due to the special locations, such as viaducts, cities, canyons, and tunnels. Then navigation and positioning cannot be achieved. Other sensors must be used to collect location information, including Wi-Fi, Bluetooth, ultra-wideband, inertial measurement unit (IMU) sensors, etc (Brena et al., 2017). In the assisted positioning systems, the inertial navigation system (INS) has been widely studied and applied due to the signal range, stability, and cost, in which the IMU is the primary sensor (Liu et al., 2021).

The INS simultaneously measures the carrier motion's angular velocity and linear acceleration by gyroscope and accelerometer. Then it solves the real-time navigation information such as 3D attitude, velocity, and carrier position (Poulose and Han, 2019; Chen et al., 2021a). The INS has essential features such as comprehensive information, a fully autonomous mechanism, and easy realization. The INS can work continuously and stably under various environmental disturbances (Soni and Trapasiya, 2021). The INS first measures the angular velocity information of the carrier by its gyroscope and further calculates the attitude information of the airline. Then the attitude information is used to support the decomposition of the accelerometer measurements. Finally, the detailed navigation information of the carrier is obtained by the transformation of the carrier coordinate system into the navigation coordinate system and then performing the navigation calculation (Cheng et al., 2022). However, the 3D attitude, velocity, and position solved in real-time in INS are achieved by primary and secondary integration of inertial data. The result of such operations will be the measurement error and noise error at the initial operation time, which will be amplified as the operation time increases, eventually leading to an increase in the position error. For this reason, deep learning methods can be considered to avoid generating cumulative errors. Only inertial measurement data and real position information are deemed for end-to-end network training. The accurate estimated position information can be obtained based on the measurement.

With machine learning and deep learning development, various networks are used for navigation and positioning. However, deep networks have different structures and parameters, which treat different data with different effects. Moreover, various situations in life will generate temporal data with different characteristics and differences in the characteristics of different data. Solving the location estimation problem with only one type of deep network is chanllenging. In order to reflect the data characteristics in various modes, the sample entropy is chosen as the metric of the time series complexity. It measures the time series complexity and the probability of generating a new mode when the dimensionality changes. The greater the generating probability, the higher the complexity degree, and the greater the entropy value. The standard deviation and sample entropy of various motion modes are shown in Table 1. It shows that the data presents distinguishing features in different modes. Therefore, this paper focuses on the positioning problem with feature matching and deep learning. We only use the motion data collected by the INS to estimate the position information. Different deep networks are studied and selected according to the data features to avoid the cumulative error problem of traditional inertial navigation. Better positioning accuracy can be achieved in various situations. Firstly, the inertial measurement data are fed into wavelet and one-sided Fourier transform for feature extraction. Secondly, the extracted data are classified by dynamic time regularization and nearest neighbor algorithm. Finally, according to the data class, the data are fed into the matched deep network for position estimation.

The rest of this paper is organized as follows. Section 2 describes the existing methods for location estimation. Section 3 describes the proposed location estimation method based on feature mode matching with a deep network model in detail in this paper. Section 4 conducts related experiments on the Oxford inertial mileage dataset and discusses the results. Section 5 concludes the paper and the directions for future research.

2. Related works

2.1. Traditional navigation and positioning methods

Traditional navigation and positioning techniques are mainly divided into two types: position determination and track projection. Among them, the position determination method relies on the external known position information for positionings, such as satellite navigation, astronomical navigation, and matching navigation. The voyage position projection method is a method to project the following instantaneous position information by measuring the bearing and distance information of the carrier movement under the condition that the initial instaneous position information is known, such as inertial navigation, magnetic compass, and odometry (Duan, 2019). Satellite and inertial navigation are still the most familiar navigation methods to the public at this stage. They are the most widely used, studied, and intensively researched navigation and positioning methods.

IMU sensors used as portable navigation applications in navigation and positioning generally have the characteristics of negligible mass, small size, low cost, and low power consumption (Huang, 2012). However, IMUs have poor performance, and it will be challenging to meet the navigation and positioning needs if they are not limited. The IMU-based positioning technique includes two solutions: the pedestrian dead reckoning (PDR) and the strap-down inertial navigation system (SINS). The PDR is based on step length estimation, which limits the propagation of inertial guidance errors through constrained models such as zero velocity correction. In the literature (Skog et al., 2013), the heading error of the PDR system is effectively eliminated by installing the IMU-based PDR positioning system on both feet. It uses the maximum distance between the two feet to constrain the positioning result of the PDR system. In the literature (Foxlinejicgs, 2005), an indoor pedestrian inertial navigation and positioning system on foot has been proposed. The inertial navigation algorithm divides the pedestrian's gait into zero velocity and motion phases. It reduces the positioning error by estimating and suppressing the inertial sensor error in the zero-velocity interval (Zheng et al., 2016). This algorithm has stability and high accuracy advantages (Zhang et al., 2018). The two solution methods have different principles and advantages.

2.2. Positioning methods with IMU

The current navigation method is multi-sensor fusion. A combined GNSS/INS navigation and positioning method is

TABLE 1 Standard deviation and sample entropy of data in different motion modes.

Mode	Handbag	Handheld	Pocket	Running	Slow walking	Trolley
Standard deviation/m	0.9409	1.1345	1.0581	0.9190	1.2245	1.1431
Sample entropy	2.0050	1.9365	2.0394	2.1837	2.1466	2.1048

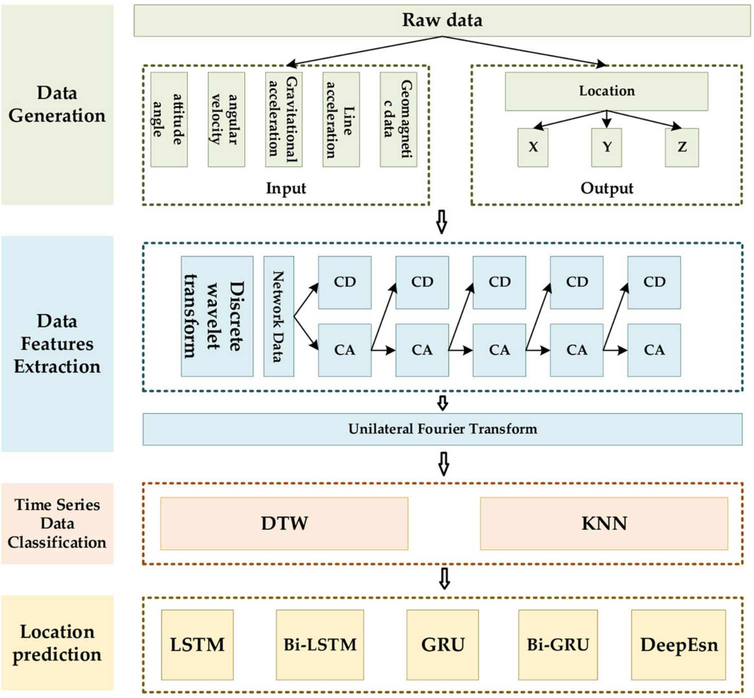


FIGURE 1
The framework of the location estimation is based on deep network matching mode features.

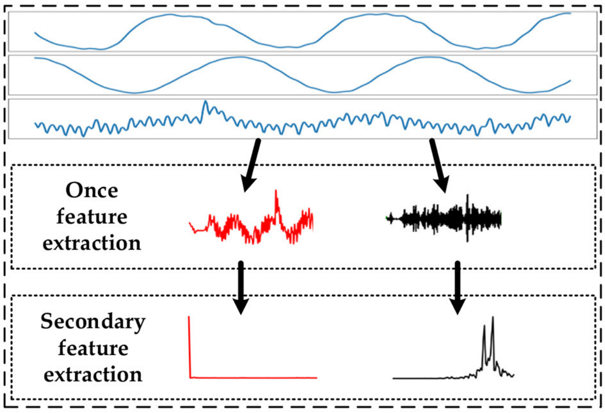


FIGURE 2
Schematic diagram of data feature extraction.

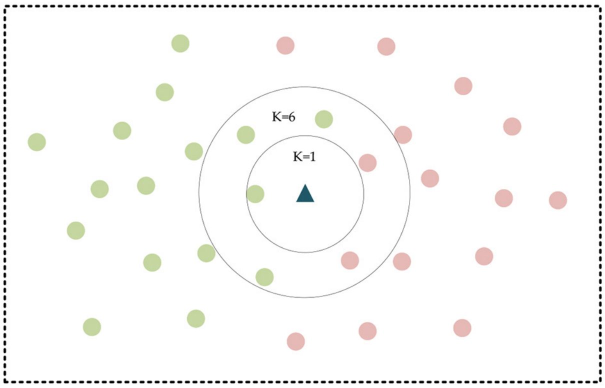


FIGURE 3
KNN schematic diagram.

proposed for pedestrian navigation with poor robustness of positioning accuracy and discontinuous position coordinates in indoor and outdoor environments (Wang, 2018; Zhu et al., 2018). In the literature (Liu et al., 2017), Wireless Sensor Networks (WSN) were fused with INS using Kalman to correct the error of firefighters in the forestry field. The advantages of the combined navigation system are reflected in the autonomous inertial navigation when there is no signal from GNSS to ensure the continuity of navigation and the combined navigation when there is a GNSS signal to ensure the navigation accuracy by GNSS constraining the error of INS.

Theoretical studies and experimental validation have been carried out for the filtering methods of GNSS/IMU combined navigation systems. More non-linear filtering algorithms have been proposed successively. The extended Kalman filter algorithm for model error prediction is applied to GNSS/INS combined navigation (Jin et al., 2023). The trace-free Kalman filter algorithm with constrained residuals fuses GPS and PDR positioning information, effectively suppressing the cumulative heading error drift (Niu and Lian, 2017). Particle filtering and robust filtering algorithms can improve the combined navigation filtering algorithm. The information from inertial navigation is fused using particle filtering to improve indoor positioning accuracy (Masiero et al., 2014). A volumetric Kalman filtering algorithm based on gated recurrent unit (GRU) networks has been proposed in the literature (Wang et al., 2022a). The filter innovations, prediction errors, and gains obtained from the filter are used as inputs to the GRU network, and the filter error values are used as outputs to train the network. End-to-end online learning is performed using the designed fully connected network, and the current state of the target is predicted. In Li et al. (2020), a hybrid algorithm based on the GRU and a robust volume Kalman filter is proposed to achieve a combined INS/GPS. It can provide high-accuracy positioning results even when GPS is interrupted. In the literature (Gao et al., 2020), an adaptive Kalman filter navigation algorithm is proposed that adaptively estimates the process noise covariance matrix using reinforcement learning methods. A sideslip angle estimation method combining a deep neural network and a non-linear Kalman filter has been proposed in the literature (Kim et al., 2020). The estimation of the deep neural network is used as a new measure of the non-linear Kalman filter, and its uncertainty is used to construct the adaptive measurement covariance matrix. The effectiveness of the algorithm is verified by simulation and experiment. According to the actual engineering requirements, when one of the system's subsystems does not work, this subsystem is removed in the fusion process, which improves the system's stability and is fully applied in various practical projects. The combined navigation technology mainly uses the positioning characteristics of INS and GNSS to combine them effectively and take advantage of their respective advantages to accomplish navigation tasks (Wu et al., 2020).

However, when GNSS is affected by the external environment, its poor anti-jamming capability makes it impossible to properly combine GNSS and INS technologies for navigation, and only INS technologies combined with depth networks can be relied on for navigation and positioning to compensate for the lack of GNSS. The literature (Yang, 2019) divides the positioning process into offline and online. In the offline process, the DNN model is trained using the signals from the signal towers, while in the online phase, the positioning process is implemented using the existing model. The literature (Wang, 2019) converts the visual information into one-dimensional landmark features using a convolutional neural network (CNN) based landmark detection model. In contrast, the wireless signal features are extracted using a weighted extraction model, and finally, the position coordinates are estimated using a regression method. The literature (Cheng et al., 2021) considers the continuity of wireless signals in the time domain during localization. It uses long short-term memory (LSTM) and temporal convolutional network (TCN) to extract features from signal sequences and calculate the localized object's position. A new AI-assisted approach for integrating high-precision INS/GNSS navigation systems is proposed in the literature (Zhao et al., 2022). Position increments during GPS interruptions are predicted by CNN-GRU, where CNN extracts multidimensional sequence features rapidly, and GRU models the time series for accurate positioning. In the literature (Liu et al., 2022), a GPS/INS neural network (GI-NN) is proposed to assist INS. The GI-NN combines CNN and GRU to extract spatial features from IMU signals and track their temporal features to build a relational model and perform a dynamic estimation of the vehicle using current and past IMU data. This paper will focus on the different effects of different networks when dealing with different data sets, and the adapted networks can improve the localization accuracy of the corresponding data.

Accurate positioning is difficult to achieve in complex environments, and the fusion of multiple technologies will solve this challenge. Neural networks are begin to significantly impact inertial navigation, where data feature analysis has been a vital issue.

3. Location estimation based on feature mode matching with deep network models

3.1. Estimation framework of feature extraction and deep networks

The data feature should be extracted first for accurate location prediction for various motion modes. We use the discrete wavelet, and Fourier transforms for different data, then classify and identify the extracted features, and finally select the deep network models that are compatible with the features. The networks are selected from the typical LSTM, bi-directional long short-term memory (Bi-LSTM), GRU, bi-directional gated recurrent unit (Bi-GRU), and deep echo state network (DeepEsn) networks. The structure of the location adaptive estimation method for the automatic matching of deep networks is shown in Figure 1.

3.2. Feature extraction and classification

Time-series data are recorded in chronological order over a specified period. All data in the same data column are of the same caliber and are comparable (Wang et al., 2021). Since time-series data are usually accurate records of system information, they reflect the trend of system changes over time by describing the state of things or phenomena, which often implies the potential laws and characteristics of the system (Kong, J. et al., 2023). Therefore, uncovering and exploiting these laws and characteristics through studying time series data is an effective means of bringing the value of time series data into play (Kong, J.-L. et al., 2023). It is also possible to classify the time series data by comparing the laws and characteristics in the time series data. Different categories of time series data will correspond to different data processing methods so that the characteristics of the data can be used more effectively.

The data features are extracted by performing two sequential processes on the data using the discrete wavelet transform and the Fourier transform. The first feature extraction is a discrete wavelet transform of the time-series data, and the second feature extraction is a Fourier transform of the first extracted feature sequence. The feature extraction process is shown in Figure 2.

The wavelet transform has the properties of local variation, multiresolution, and decorrelation (Vidakovic and Lozoya, 1998). It translates the data at different scales to obtain wavelet coefficients. The discrete wavelet transform (DWT) in wavelet transform decomposes the data by high-pass and low-pass filters to produce an approximate component of approximate (CA) and component of detail (CD), respectively (Deng, 2021; Wang et al., 2022b).

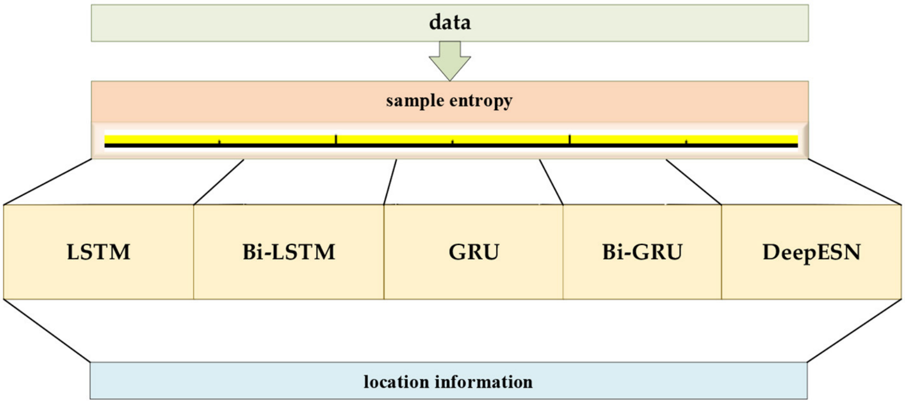


FIGURE 4
Algorithm flow of selecting networks for different mode data.

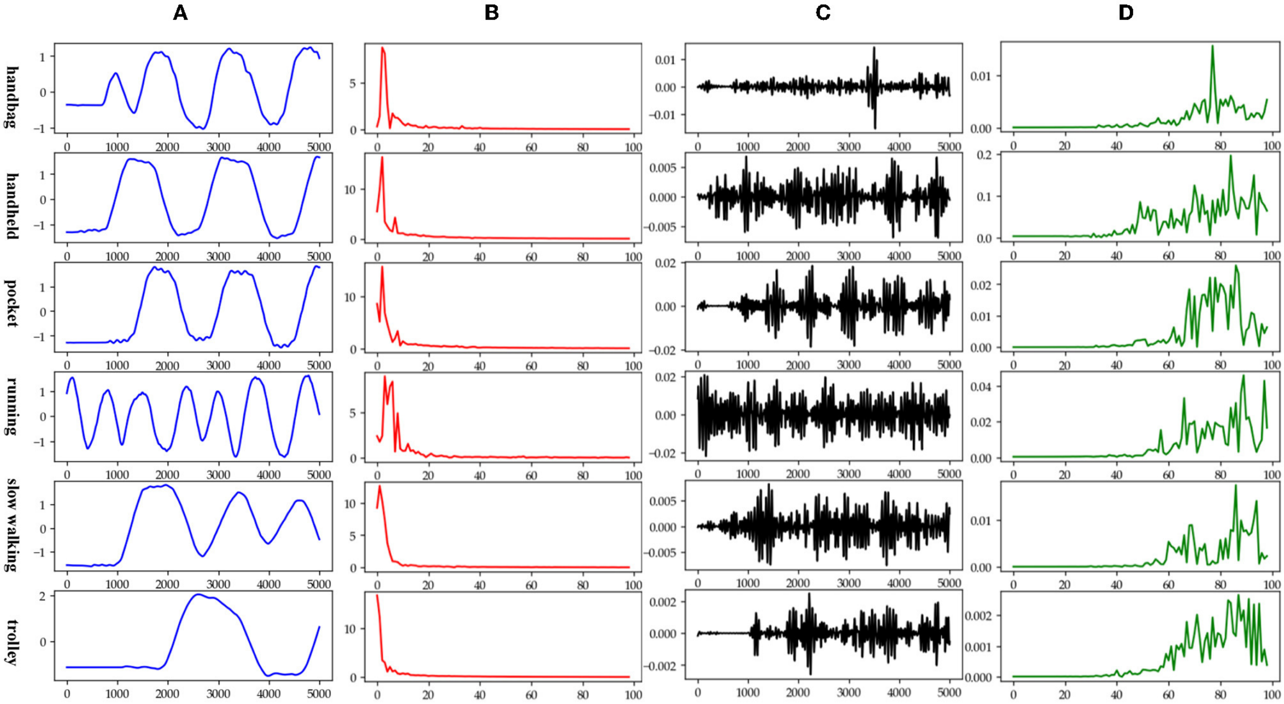


FIGURE 5
Data decomposition results for each motion mode: **(A–D)** are the unilateral Fourier variation of CA and CA after decomposition of the fifth-order discrete wavelet transform, and the unilateral Fourier variation of CD and CD after decomposition of the fifth-order discrete wavelet transform, respectively.

When performing a multi-order DWT, the CD is processed using a high-pass filter, while CA will continue to be decomposed. The multi-order decomposition is used to correct the high-frequency information in the data and effectively extract the data features. The discrete wavelet transforms equation and the DWT algorithm expressions are shown in equations (1) and (2), respectively.

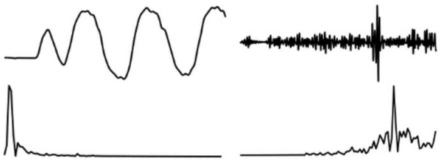
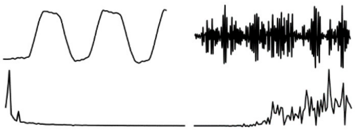
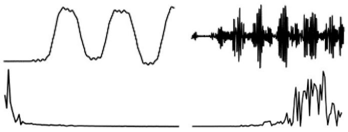
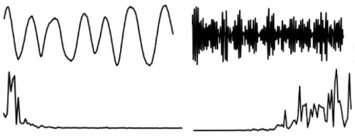
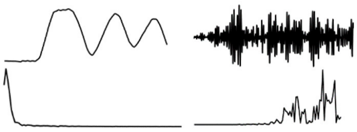
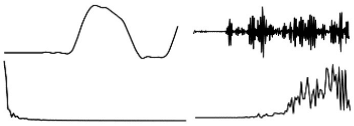
$$f(t) = \sum_{j=-\infty}^{j} \left[\sum_{k=-\infty}^{\infty} d_{j,k}\phi_{j,k}(t) + \sum_{k=-\infty}^{\infty} c_{j,k}\phi(t) \right] \quad (1)$$

where $\sum_{k=-\infty}^{\infty} c_{j,k}\phi(t)$ is approximate data (low-frequency). $\sum_{k=-\infty}^{\infty} d_{j,k}\phi_{j,k}(t)$ is the detail data(high-frequency). $\phi_{j,k}(t)$ is the basic wavelet function. $\phi(t)$ is the scale function.

$$A_i f(t) = A_{i+1} f(t) + D_{i+1} f(t) \quad (2)$$

where $A_i f(t)$ is the low-frequency part of the wavelet decomposition of the first layer. $A_{i+1} f(t)$ and $D_{i+1} f(t)$ are the

TABLE 2 Network selection regarding the feature mode matching.

Mode category	Sample entropy	Decomposition results graph	Deep network to be selected	Remarks
Handbag	2.005		Bi-LSTM GRU	The data has a clear cyclical trend, with a dense and small magnitude of detailed trends.
Handheld	1.936		DeepEsn LSTM	The data has a clear cyclical trend, and the detailed features are intensive and cyclical.
Pocket	2.309		DeepEsn	The data has a clear cyclical trend, and the detailed features are intensive and cyclical.
Running	2.183		LSTM GRU	The data has a clear cyclical trend and segmented period, with intensive and detailed features.
Slow walking	2.146		DeepEsn Bi-GRU	The data has a clear cyclical trend, and the detailed features are intensive and cyclical.
Trolley	2.104		Bi-GRU	The data has a clear and long cyclical trend; sudden changes dominate the detailed features.

low frequency part and high frequency part of the next layer of decomposition, respectively.

The Fourier transform is a standard method for analyzing signals. The process converts a continuous signal that is non-periodic in the time domain into a continuous signal that is non-periodic in the frequency domain. The same principle can be used to analyze and process time-series data, and the Fourier transform is shown in equation (3).

$$\mathrm{F}(\omega) = \mathrm{F}\left(\mathrm{f}(\mathrm{t})\right) = \int_{-\infty}^{+\infty} \mathrm{f}(\mathrm{t})\, \mathrm{e}^{-\mathrm{i}\omega \mathrm{t}} \mathrm{dt} \qquad (3)$$

where $\mathrm{F}(\omega)$ is the image function of $\mathrm{f}(\mathrm{t})$. $\mathrm{f}(\mathrm{t})$ is the original image function of $\mathrm{F}(\omega)$.

Classification of temporal data is mainly divided into benchmark methods, which use feature similarity as a determination. Traditional methods classify data by underlying modes and features, and deep learning classification methods.

TABLE 3 Data characteristics of each motion mode.

Mode	p_{mean}	$p_{variance}$	$p_{peakedness}$	$p_{skewness}$
Handbag	0.4302	1.59	33.633	5.619
Handheld	0.7691	4.04	40.954	6.008
Pocket	0.8235	3.98	33.098	5.292
Running	0.5978	2.66	16.142	4.015
Slow walking	0.6053	3.99	22.027	4.672
Trolley	0.6772	4.99	34.816	5.727

Classification by deep learning methods performs very well on image, audio, and text data and can quickly update data using batch propagation (Jonathan et al., 2020). However, they are unsuitable as general-purpose algorithms because they require large amounts of data. Classical machine-learning problems are usually better than tree collections. Moreover, they are computationally intensive during training and require more expertise to tune the parameters. The Oxford Inertial Mileage dataset is characterized by various types of time-series data and a small number of data sequences. Compared to deep learning methods that require architecture and hyperparameter tuning, traditional methods that determine the similarity of classification features are more straightforward and faster and can achieve good classification results.

It usually classifies features using Euclidean distance and dynamic time warping (DTW). They are set as the similarity measure by calculating the distance between the original or temporal data after feature representation (Pimpalkhute et al., 2021). Then it uses the nearest neighbor classifier for classification. This similarity metric-based method for classifying temporal data is simple in principle and structure, easy to implement, and is considered the benchmark method for classifying temporal data. The K-Nearest Neighbors (KNN) algorithm has been the simple and typical classification algorithm. In KNN, when a new value *x* is predicted, the class to which *x* belongs is determined based on its class from the nearest K points (Zhang, 2021). The KNN schematic is shown in Figure 3, in which the green and red dots represent the two categories, and the triangular points are the points to be classified.

The distance calculation is usually chosen as the Euclidean distance with the equation.

$$d = \sqrt{\left(x_1 - y_1\right)^2 + \left(x_2 - y_2\right)^2 + \cdots + \left(x_n - y_n\right)^2} = \sqrt{\sum_{i}^{n} \left(x_i - y_i\right)^2} \tag{4}$$

where *x* and *y* are coordinates in the two-dimensional plane, and the subscripts are the ordinal numbers of the data points.

The dynamic time regularization algorithm is a proposed metric between sequences for time-series data. The DTW algorithm finds the best correspondence between two observed sequences by regularizing the time dimension with certain constraints. Therefore, DTW is suitable for classifying sequences with different frequencies or phases. DTW uses the idea of dynamic programming to calculate the optimal path between two sequences, where the dynamic transfer equation is as follows.

$$D(i,j) = Dist(i,j) + \min[D(i-1,j), D(i-1,j), D(i-1,j)] \tag{5}$$

where $D\left(i,j\right)$ is the coordinate of the distance matrix, $Dist\left(i,j\right)$ is the calculated Euclidean distance, and $D\left(i-1,j\right)$, $D\left(i,j-1\right)$, and $D\left(i-1,j-1\right)$ are the lower left 3 elements of $D\left(i,j\right)$, respectively.

3.3. Feature model matching with deep networks

The data in different modes can be selected to fit the depth network according to their different feature information. From parts 3.1 and 3.2 of this paper, different mode features can be extracted and distinguished effectively, then according to the different mode features to match the depth network that fits with the features, which can better perform the performance of network estimation. By mathematically analyzing the time-series data features, the features can be characterized by the mean, variance, skewness, and kurtosis of the data features. The mean and variance can reflect the overall trend in the data set, and the skewness and kurtosis can reflect the local details in the data set. Equation (6) shows the overall form of the evaluation data features, f_a is the evaluation in a mode which contains p_{mean}, $p_{variance}$, $p_{peakedness}$ and $p_{skewness}$, representing the mean, variance, kurtosis, and skewness of the feature data, respectively.

$$f_a = f\left(p_{mean}, p_{variance}, p_{peakedness}, p_{skewness}\right) \tag{6}$$

The networks selected in this paper include LSTM, Bi-LSTM, GRU, Bi-GRU, and DeepEsn. LSTM networks consist of forgetting, input, and output gates, which can handle longer data sequences and solve the problems of gradient disappearance and gradient explosion problems. The GRU network is a suitable variant of the LSTM network, and its structure is more concise than the LSTM. The two-way network can correlate the next-moment state information and the previous-moment state information to estimate the output with the previous and future states. Finally, as an improved ESN network, the DeepEsn develops from a single reserve pool to a deep learning network consisting of a multi-layer reserve pool structure in series. The characteristics of leaky integral-type neurons in each reserve pool will effectively improve the memory of network history information. The networks are selected according to the trend information and detailed information of the data features according to the characteristics of the five deep network models selected in this paper. The network selection of data modes is shown in Table 2.

The basis for selecting networks for different modes is based on different temporal complexity between data to reflect the data characteristics of different modes to determine which network should be selected. The temporal complexity can be judged based on the size of the sample entropy to determine the network for location estimation and achieve location estimation. The sample entropy has a direct relationship with the complexity of the data. The higher the sample entropy, the higher the complexity of the data, and the more the deep network with higher processing data is needed to process to achieve better results. The deep networks selected in this paper contain LSTM, GRU, Bi-LSTM, Bi-GRU, DeepESN networks. GRU is an improved network of LSTM network, but the prediction ability of the network is similar, only in the training speed of the model is improved; Bi-LSTM and Bi-GRU networks as improved networks of LSTM, GRU, the network model from only Bi-LSTM and Bi-GRU networks are improved networks of LSTM and GRU. The network models use information from the forward direction to the forward and backward movement, which makes the network models process more complex data and improve prediction accuracy; DeepESN networks have more complex layers than other network models and can handle more complex data. Therefore, the selection of the deep network is directly related to the complexity of the data. The higher the complexity of the data, the

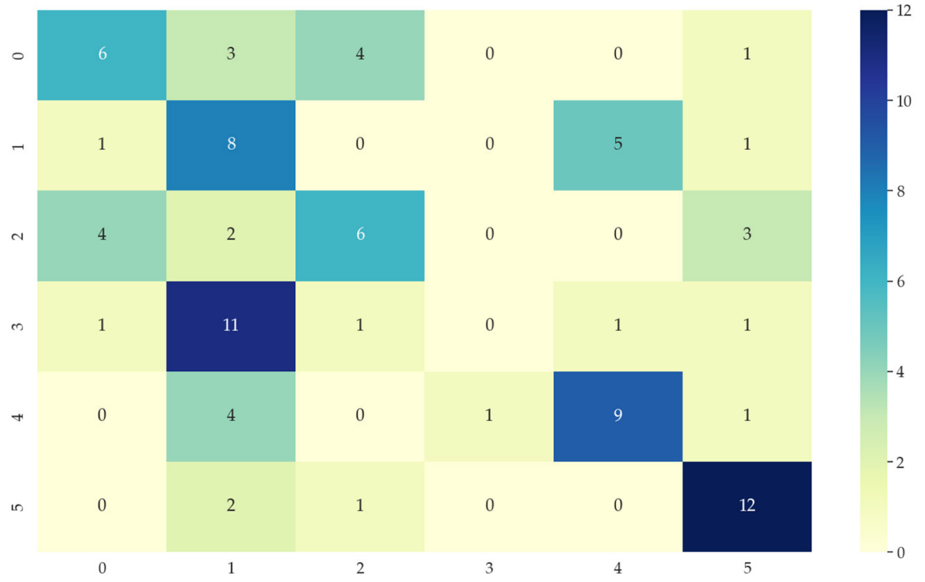


FIGURE 6
Classification confusion matrix from DTW + KNN algorithm of original data.

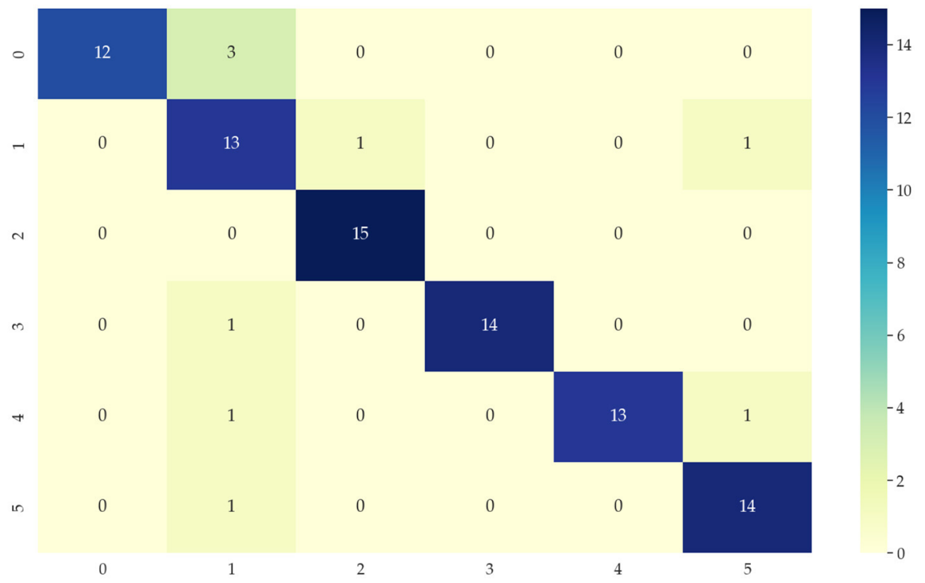


FIGURE 7
Classification confusion matrix after data feature extraction.

higher the sample entropy, and the deep network should be more complex. The flow chart of network selection for different modes is shown in Figure 4.

Figure 4 shows the algorithm flow of selecting networks for different data modes. The original data are first subjected to the calculation of sample entropy. The corresponding network is chosen according to the magnitude of the sample entropy, and finally, the selected position estimation is achieved using the selected depth network.

4. Experiment and result

In this section, we use the Oxford Inertial Ranging Dataset (OxIOD), classify the data based on its features, and select compatible networks from LSTM, Bi-LSTM, GRU, Bi-GRU, and DeepEsn network models based on various types of features. The PDR is also set as the baseline model, which uses the movement speed and forward direction to infer the positioning process. Finally, extensive experiments are conducted to verify

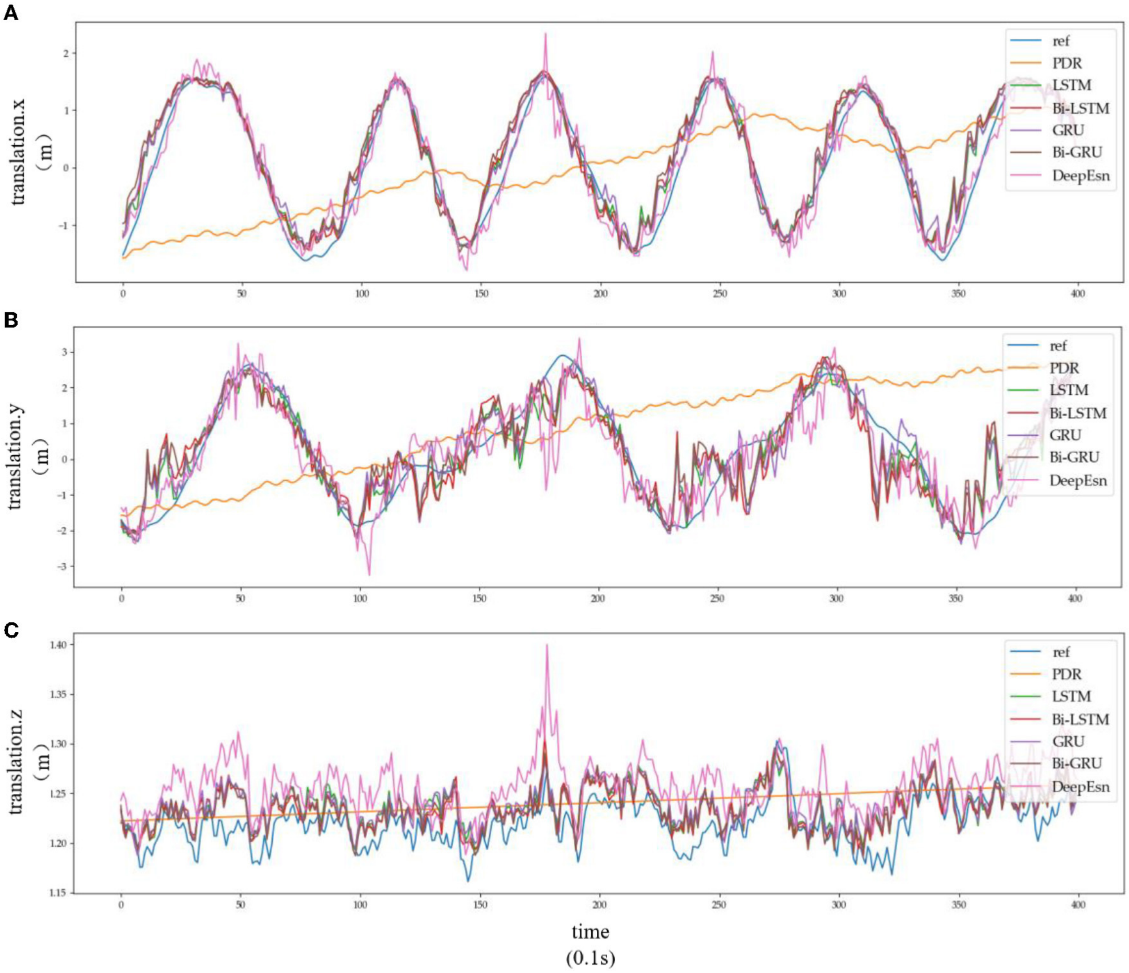


FIGURE 8
Location estimation results of different deep networks in running mode: **(A–C)** is the Location estimation results of translation.x, translation.y, and translation.z in running mode, respectively.

the appropriateness of the network selection from the estimated network results.

4.1. Data sets and experiment setting

In this paper, we use the OxIOD dataset, in which ground truth data for indoor walking is collected using the Vicon optical motion capture system, which is known for its high accuracy (0.01 m in position and 0.1 degrees in direction) in target localization and tracking (Chen et al., 2020; Kim et al., 2021). The IMU sensors on smartphones. It includes data from four off-the-shelf consumer phones and five different users and data from different locations and motion states of the same pedestrian, including handheld, pocket, handbag, and stroller data in a normal walking motion, slow walking, and running (Markus et al., 2008; Chen et al., 2021b). The raw inertial measurements were segmented into sequences with a window size of 200 frames (2 s) and a step length of 10 frames. OxIOD's data is extensive and has highly accurate actual values, making it suitable for deep learning methods. At the same time, the dataset contains a wide range of human movements that can represent everyday conditions, providing greater diversity.

For each type of data, different divisions were performed. The training set is 7 sequences for handbag data, 10 sequences for pocket data, 20 sequences for handheld data, 6 sequences for running data, 7 sequences for slow walking data, and 12 sequences for cart data. The test set is the rest of the sequences. The input of each experiment below is 15 data items of sensors in the dataset, and the output is 13 data items, namely, changing displacement, heading angle, changing heading angle, average speed, speed of heading angle, changing the speed of heading angle, translation.x, translation.y, translation.z, rotation.x, rotation.y, rotation.z, and rotation.w. This paper will focus on the output position information translation.x and translation.y for experimental study.

4.2. Feature extraction and classification

The data in various modes have different data characteristics. We can qualitatively find the individual characteristics and assign

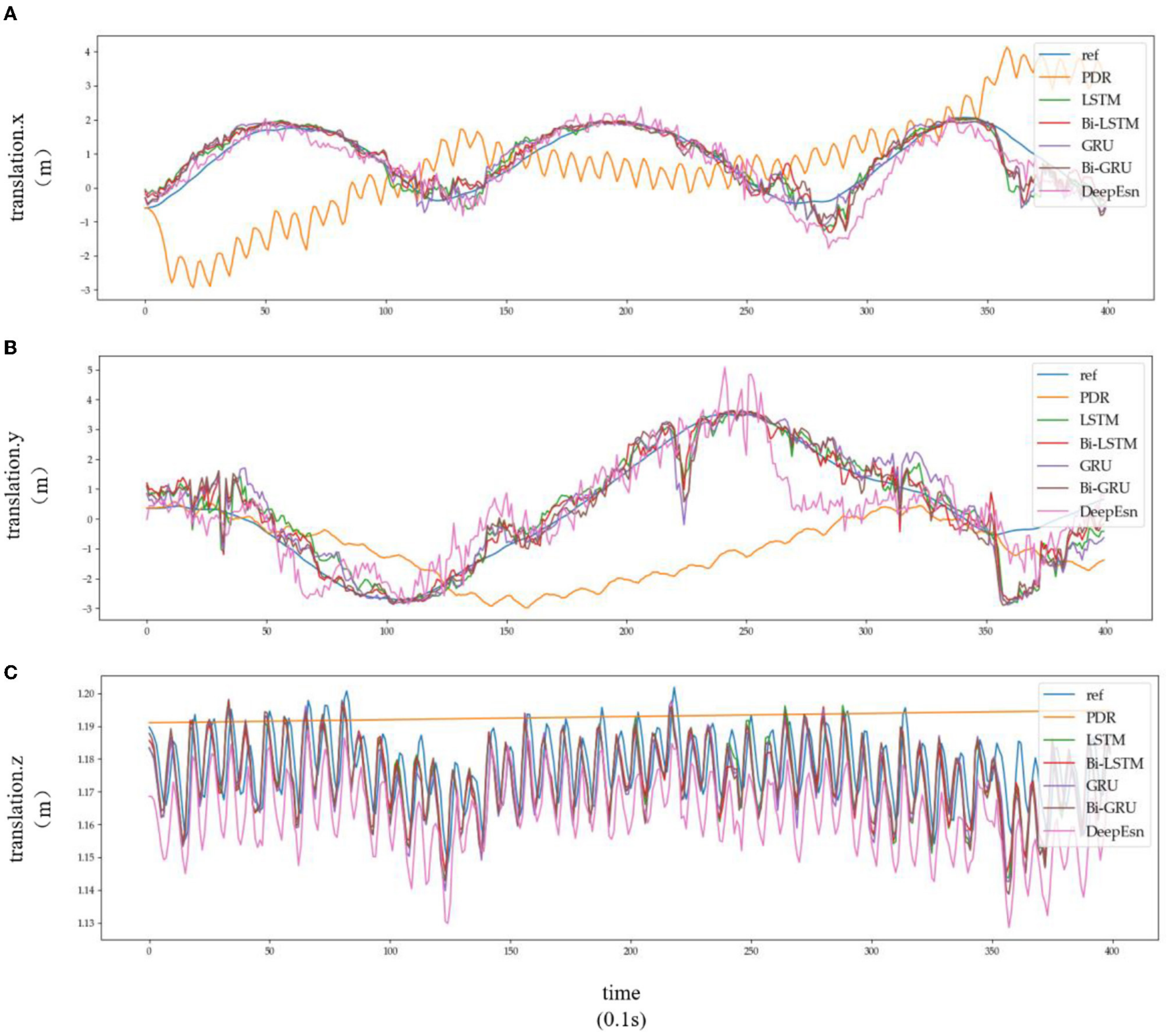


FIGURE 9
Location estimation results of different deep networks in slow walking mode: **(A–C)** is the Location estimation results of translation.x, translation.y, and translation.z in slow walking mode, respectively.

the appropriate deep network model through data decomposition and feature extraction. Wavelet decomposition can decompose signals at different scales, and the choice of different scales can be determined according to different objectives. Wavelet decomposition achieves feature extraction by decomposing the low-frequency and high-frequency features of the data. In this paper, the db5 wavelet, which is widely used and has a better processing effect, is chosen as the wavelet base, and the number of decomposition layers is chosen as 5. The decomposition results of the data of various modes are shown in Figure 5.

The subplot in Figure 5 represents the results of the decomposition of handbag, handheld, pocket, running, slow walking, and trolley mode data in each of the 6 rows of subplots from top to bottom. The figure's four subplots from left to right are columns A, B, C, and D. The blue and black line subplots in columns A and C of the figure show the CA and CD after decomposition of the fifth-order discrete wavelet transform, which has a data volume of 5,000, allowing differences in frequency, amplitude, and other relevant information to be observed. However, the results cannot be directly observed quantitatively. The figure's red and green line plots in columns B and D are the

TABLE 4 Evaluation metric of EvaMe in the running and slow walking mode data.

Mode	PDR	LSTM	Bi-LSTM	GRU	Bi-GRU	DeepEsn
Running	0.89214	**0.25351**	0.28212	0.27012	0.30051	0.36852
Slow walking	2.12199	0.30596	0.30891	0.3048	**0.29656**	0.446

The bold values indicate the best evaluation metrics for the various methods in running and slow walking mode, respectively.

TABLE 5 Evaluation metrics for each network in the running and slow walking modes.

Mode			PDR	LSTM	Bi-LSTM	GRU	Bi-GRU	DeepEsn
Running	translation.x	RMSE	1.227133	0.365245	0.375894	0.369127	0.423125	0.333949
		MSE	1.505855	0.133404	0.141296	0.136255	0.179035	0.111522
		R	0.077798	0.955243	0.949109	0.950859	0.941615	0.941075
		R2	−1.836009	0.857052	0.848594	0.853997	0.808156	0.880499
	translation.y	RMSE	2.003978	0.735532	0.822200	0.790001	0.849290	0.813939
		MSE	4.015928	0.541007	0.676013	0.624102	0.721294	0.662496
		R	0.086135	0.874320	0.840365	0.856647	0.829742	0.847130
		R2	−1.339996	0.761948	0.702543	0.725385	0.682619	0.708490
	translation.z	RMSE	0.030548	0.023053	0.022737	0.023073	0.022583	0.040435
		MSE	0.000933	0.000531	0.000517	0.000532	0.000510	0.001635
		R	0.324380	0.803936	0.795007	0.811166	0.804379	0.709903
		R2	−7.316397	0.503342	0.516893	0.502510	0.523396	−0.527917
Slow walking	translation.x	RMSE	1.770301	0.499401	0.511289	0.498108	0.449248	0.515972
		MSE	3.133964	0.249401	0.261417	0.248111	0.201824	0.266228
		R	0.011000	0.921557	0.925347	0.917857	0.928084	0.906085
		R2	−0.268839	0.808397	0.799166	0.809388	0.844949	0.795470
	translation.y	RMSE	2.298411	1.000857	1.000378	0.990982	1.001029	1.178784
		MSE	5.282694	1.001714	1.000756	0.982046	1.002059	1.389532
		R	0.153591	0.861649	0.862975	0.863281	0.862232	0.807027
		R2	−4.294373	0.726944	0.727205	0.732305	0.726850	0.621229
	translation.z	RMSE	0.018717	0.005231	0.005327	0.005518	0.005566	0.013099
		MSE	0.000350	0.000027	0.000028	0.000030	0.000031	0.000172
		R	−0.130370	0.921210	0.917341	0.916645	0.913296	0.891483
		R2	−287.7802	0.845402	0.839658	0.827948	0.824970	0.030516

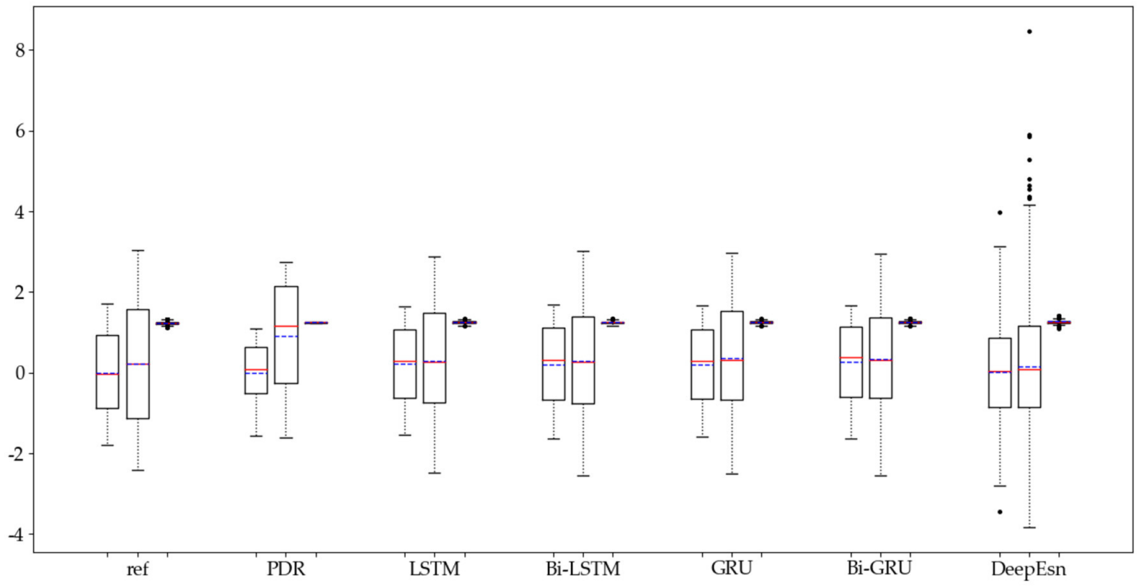


FIGURE 10
Distribution of location estimation results for different networks in the running mode.

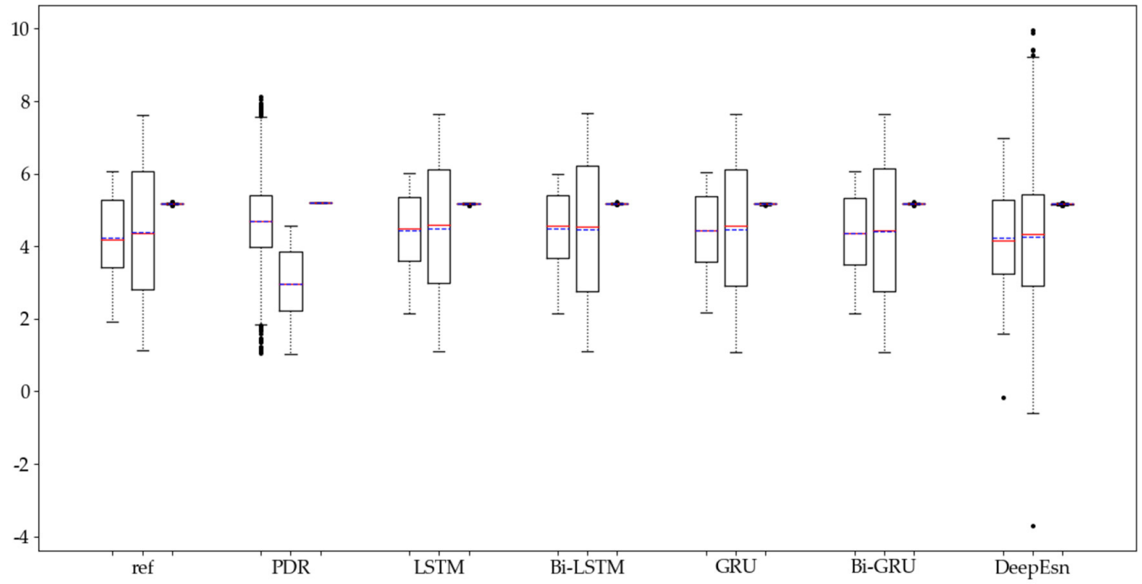


FIGURE 11
Distribution of location estimation results for different networks in the slow walking mode.

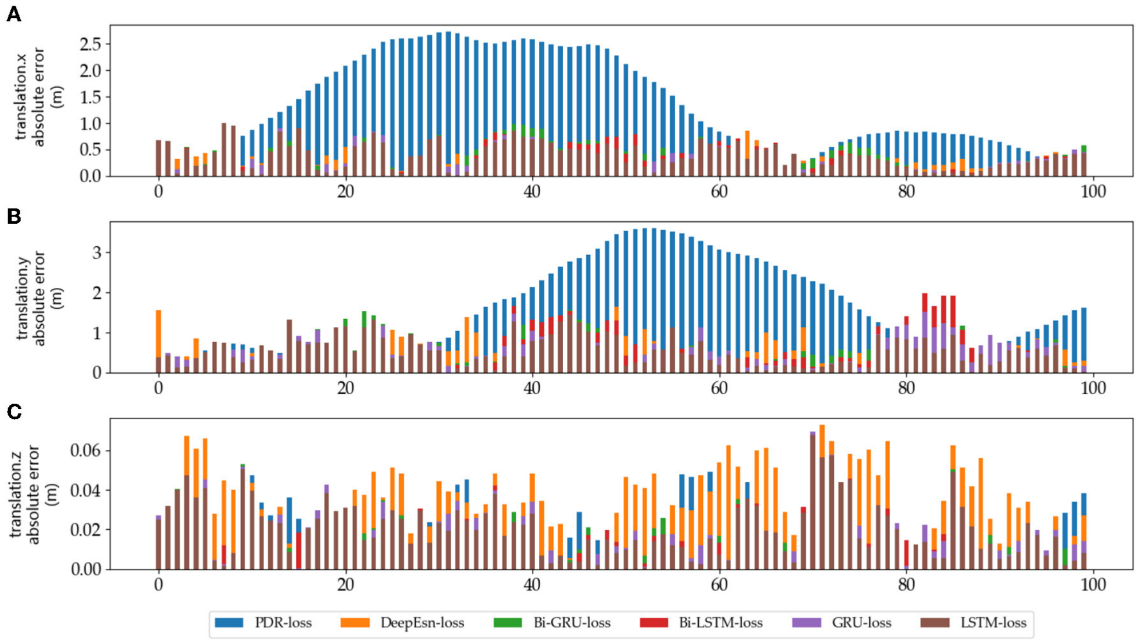


FIGURE 12
The absolute errors of different methods in running mode: **(A–C)** is the absolute errors of translation.x, translation.y, and translation.z in running mode, respectively.

spectrum plots of CA and CD with unilateral Fourier variation, respectively. Based on the spectrum plots of CA and CD, various modes can be effectively distinguished, and the corresponding deep network model. The Bi-LSTM network is selected for the data in the handbag mode, while the Bi-LSTM network is selected for the data in the handheld, pocket, running, or slow walking modes. The DeepEsn, LSTM, and Bi-GRU networks are selected for position estimation for the trolley mode. The number of reserve layer layers in the DeepEsn network and the number of neurons in the reserve layers in the DeepEsn network are chosen separately according to the situation. The results of the data feature representation of each mode are shown in Table 3.

Identifying data types for classification means that the input temporal data is used to correctly distinguish which category of the six modes mentioned in 4.1 is identified to correctly select the appropriate network model. This paper chooses the KNN algorithm of the dynamic time regularization (DTW) algorithm for classification. However, the direct recognition and classification

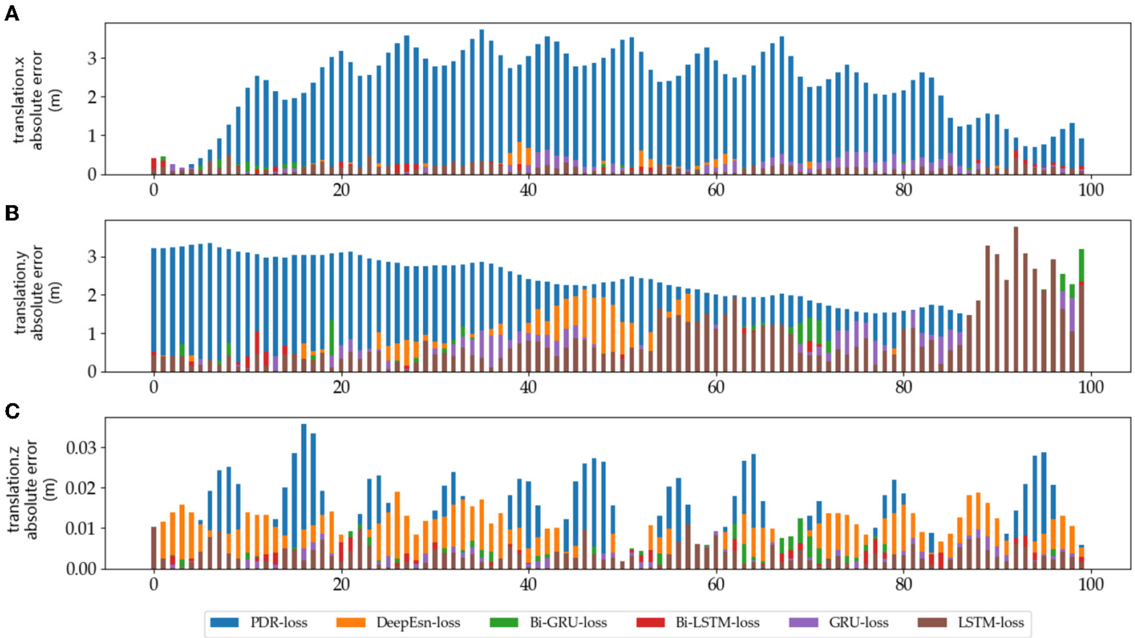


FIGURE 13
The absolute errors of different methods in slow walking mode: **(A–C)** is the absolute errors of translation.x, translation.y, and translation.z in slow walking mode, respectively.

of the original data do not extract the hidden features in the data well, and its classification results are poor, as shown in Figure 6. Therefore, we need to identify and classify the sequences after extracting the features to improve the accuracy. After decomposing the features extracted by the 4.2 part of the data, the accuracy of the classification results can reach 90%, and the classification results are shown in Figure 7. 0, 1, 2, 3, 4, and 5 in the horizontal and vertical coordinates correspond to 6 types of mode data.

4.3. Location estimation in different modes

This paper uses the OxIOD dataset and attempts to better solve the pedestrian inertial navigation problem using different deep neural networks depending on the model. This paper uses LSTM, Bi-LSTM, GRU, and Bi-GRU network models with two-layer networks with 128 and 64-dimensional hidden states, respectively. In contrast, DeepEsn networks use the best network structure with the number of reserve layers ranging from 1 to 7 and the number of neurons in the reserved layer ranging from 500 to 750. The models were trained using detailed split training sets for the four attachment categories mentioned above, namely, handheld (20 sequences), pocket (10 sequences), handbag (7 sequences), cart (12 sequences), running (6 sequences), and slow walking (7 sequences). The different split types of datasets were put into the neural network for training, and the input data remained IMU sensor data. The output results were selected with the location based translation.x, translation. y and translation.z data for viewing and comparison with the baseline model; the results are shown in Figures 8, 9, and Table 4. The detailed RMSE, MSE, R, and R2 evaluation metrics are shown in Table 5. Figures 8, 9 show the estimation results of the three-way location coordinates over time for the predictions of different networks on running and slow walking data, respectively, in terms of position information. The LSTM, Bi-LSTM, GRU, and Bi-GRU networks in the article experiments are all 2-layer structures, and the number of neurons per layer is 32. For the DeepESN network, the number of reserve pool neurons ranges from 400 to 600, and the number of layers ranges from 1 to 7, and the optimal result is chosen as the final structure of DeepESN.

TABLE 6 Evaluation metric of EvaMe for each method in different modes.

Mode	PDR	LSTM	Bi-LSTM	GRU	Bi-GRU	DeepEsn
Handbag	3.85124	0.22663	**0.22525**	0.22822	0.23919	0.23812
Handheld	28.9174	1.13908	1.14954	1.10937	1.14782	**0.9371**
Pocket	49.0785	0.41175	0.4226	0.3961	0.39225	**0.29352**
Trolley	2.05988	0.33444	0.37399	0.36091	**0.33257**	0.38966

The bold values indicate the best evaluation metrics for each method in handbag, handheld, pocket, and trolley modes, respectively.

Subplots a, b, and c in Figures 8, 9 represent the comparison results of the position coordinates x, y, and z for different networks of the corresponding modes, respectively.

The evaluation indexes used in the experiments were root mean square error (RMSE), mean square error (MSE), correlation coefficient (R), and coefficient of determination (R2), and then the evaluation index EvaMe was obtained by a weighted averaging method:

$$\mathrm{EvaMe} = \frac{\sum_{i}^{n} (\alpha_i \times \mathrm{E_{loss}})}{n} \tag{7}$$

where n is the number of evaluation indicators, $\alpha_i$is the weight coefficient, E_{loss} is the evaluation indicator. Because the evaluation

TABLE 7 Network structures are selected for different mode inputs.

Mode	Running	Slow walking	Handbag	Handheld	Pocket	Trolley
Network	Bi-GRU	DeepEsn	Bi-LSTM	DeepEsn	LSTM	Bi-GRU
Layers	2	6	2	2	2	2
Number of neurons	32	500	32	500	32	32

indicators selected in this paper are the error, the smaller the error, the higher the accuracy, and for the other two indicators used in this paper correlation coefficient and coefficient of determination is the closer to 1, the higher the accuracy if we want to use equation (7), will need to the correlation coefficient and coefficient of determination for error processing.

The structures selected for the DeepEsn networks in the running and slow walking modes are a 5-layer reserve layer with 600 neurons and a 6-layer reserve layer with 500 neurons, respectively. The evaluation metric EvaMe shows that the best prediction of position information is achieved by the LSTM network under running mode data. In contrast, the Bi-GRU network achieves the best prediction of location information under slow walking mode data. The distribution of data results predicted by each of its networks is shown in Figures 10, 11, and similar location information data are translated.x, translation.y, and translation.z from left to right.

The absolute error values of the predicted data results under different networks with reference data under running and slow walking mode data are shown in Figures 12, 13. The position absolute error plots are drawn by selecting 100 sets from the test set data, and the plots are translation.x, translation.y, and translation.z absolute error data.

Similarly, different deep networks estimate the location of handbag, handheld, pocket, and trolley mode data. The evaluation metrics EvaMe of their estimation results are shown in Table 6. The DeepEsn network structures in the table are the 5-layer 500 neuron reserve layer, 2-layer 500 neuron reserve layer, 7-layer 500 neuron reserve layer, and 5-layer 500 neuron reserve layer, respectively.

In Table 6, the estimation results indicate that the Bi-LSTM network should be selected for handbag mode. DeepEsn network of 2-layer and 500-neuron reserve should be selected for handheld mode. DeepEsn network of 7-layer and 500-neuron reserve should be selected for pocket mode. Bi-GRU network should be selected for trolley mode.

5. Discussion and conclusion

This paper proposes a method based on mode features and deep network matching to achieve location estimation. Firstly, feature extraction is performed on different mode data. The data's trend and detail features of the data are effectively extracted by discrete wavelet transform. Fourier transforms, and the data selection network is distinguished using mean, variance, kurtosis, and skewness mathematical indicators. Then, the classification is then performed by K-nearest neighbor and dynamic time regularization, and the classification accuracy reaches 90 from 30%, which shows the importance and necessity of data feature extraction methods. Finally, the evaluation indices of location information estimation under different modes prove the correctness and feasibility of the location estimation based on the matching method of mode features and deep networks. In this paper, the network is selected according to the decomposed mode features. The Bi-LSTM network is selected for the handbag mode, which has the trend cycle and small density amplitude. DeepEsn network is selected for the handheld mode with trend and detail cycles. The LSTM network is selected for the pocket mode with both the trend cycle and the detail cycle. And the LSTM network is selected for the running mode decomposition with a short trend cycle. The Bi-GRU network is selected for the running mode with a short trend period and dense detail. The DeepEsn network is selected after the slow walking mode decomposition with a long trend period and dense details. The Bi-GRU network is selected after the trolley mode decomposition with a long trend period. The sample entropy is used for the complexity of various mode types of data and the model's classification into categories. Many locational estimation experiments verify the feasibility of the method. Table 7. shows the network structure that should be selected for the different mode inputs.

The existing positioning and navigation techniques mostly use multiple positioning and navigation techniques to enhance the accuracy and application range of navigation and positioning through the fusion of advantages. At the same time, in this paper, we choose different adaptive depth networks for a single navigation technique with different input data modes to achieve positioning accuracy and expand the application range. In this paper, we choose different phase-adaptive depth networks to position and expand the application range. LSTM, GRU, Bi-LSTM, Bi-GRU, and DeepESN networks are more common deep networks with simple structures and parameters and easy-to-implement prediction functions. The end-to-end training approach relies on only one model and one objective function, which can circumvent the inconsistency in training multiple modules and the deviation of the objective function. The generalization can be obtained with the learning mode to solve the error accumulation problem in the traditional location solution. The model is more general for relying only on inertial data using a deep network model to estimate the position information. Only six modes are classified in this paper, and the mode types are limited. There is still a gap between the application scope and accuracy of single-location navigation and multi-positioning navigation techniques. Future work will solve these problems by replacing the adapted depth networks with more efficient and optimized neural networks and combining them with multi-location navigation techniques to improve accuracy.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

Y-TB: conceptualization. WJ and Y-TB: methodology and writing—review and editing. WJ, Y-TB, and X-BJ: writing—original draft preparation. Y-TB, X-BJ, J-LK, and T-LS: funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding

This work was partly supported by the National Natural Science Foundation of China Nos. 62203020, 62173007, and 62006008.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Brena, R. F., García-Vázquez, J. P., and Galván-Tejada, C. E. (2017). Evolution of indoor positioning technologies: a survey. *IEEE Sens. J.* 2017, 21. doi: 10.1155/2017/2630413

Chen, C., Lu, C., Wahlström, J., Markham, A., and Trigoni, N. (2021b). Deep neural network based inertial odometry using low-cost inertial measurement units. *IEEE Trans.* 20, 1351–1364. doi: 10.1109/TMC.2019.2960780

Chen, C., Zhao, P., Lu, C., and Wang, W. (2020). Markham. deep-learning-based pedestrian inertial navigation: methods, data set, and on-device inference. *IEEE Int. Things.* 7, 4431–4441. doi: 10.1109/JIOT.2020.2966773

Chen, J., Zhou, B., Bao, S., Liu, X., Gu, Z., Li, L., et al. (2021a). A data-driven inertial navigation/bluetooth fusion algorithm for indoor localization. *IEEE Sens. J.* 22, 5288–5301. doi: 10.1109/JSEN.2021.3089516

Cheng, J., Guan, H., and Li, P. (2022). Research on polar-stabilizedd inertial navigation algorithm based on the transverse geographic coordinate system. *Navigat. Position. Time* 9, 69–75. doi: 10.19306/j.cnki.2095-8110.2022.01.008

Cheng, Y., Qu, J. Q., and Tang, W. J. (2021). Design and implementation on indoor positioning system based on LSTM. *Electron. Measure. Technol.* 44, 161–166. doi: 10.19651/j.cnki.emt.2107563

Deng, Y. F. (2021). ECG signal denoising algorithm based on particle swarm optimization and discrete wavelet transform. *STI.* 6, 9–11. doi: 10.3969/j.issn.1673-1328.2021.06.006

Dong, M. T., Chen, J. J., and Ban, J. C. (2021). *A Survey of Light and Small Inertial Navigation Systems*. CCAN2021, Xiamen. doi: 10.26914/c.cnkihy.2021.048791

Duan, R. (2019). *Research on the key technology and platform design of real-time GNSS/SINS integrated navigation system*. master degree, Wuhan University, Wuhan.

Foxlinejicgs (2005). Pedestrian tracking with shoe-mounted inertial sensors. *IEEE Comput. Graph.* 25, 38–46. doi: 10.1109/MCG.2005.140

Gao, X., Luo, H., Ning, B., Zhao, F., and Jiang, J. (2020). RL-AKF: An adaptive kalman filter navigation algorithm based on reinforcement learning for ground vehicles. *Remot. Sens.* 12, 1704–1715. doi: 10.3390/rs12111704

Huang, F. (2012). *Research on the algorithm of strapdown AHRS based on MIMU*. master degree, Shanghai Jiaotong University, Shanghai.

Jin, X.-B., Wang, Z.-Y., Gong, W.-T., Kong, J.-L., Bai, Y.-T., Su, T.-L. et al. (2023). Variational bayesian network with information interpretability filtering for air quality forecasting. Mathematics.11, 837. doi: 10.3390/MATH11040837

Jin, X.-B., Wang, Z.-Y., Kong, J.-L., Bai, Y.-T., Su, T.-L., Ma, H.-J., et al. (2023). Deep spatio-temporal graph network with self-optimization for air quality Prediction. *Entropy.* 25, 247. doi: 10.3390/E25020247

Jonathan, R., Isak, K., Leon, B., and Panagiotis, P. (2020). SMILE: a feature-based temporal abstraction framework for event-interval sequence classification. *Data. Min. Knowl. DISC* 35, 372–399. doi: 10.1007/s10618-020-00719-3

Kim, D., Min, K., Kim, H., and Huh, K. (2020). Vehicle sideslip angle estimation using deep ensemble-based adaptive Kalman filter. *MSSP* 144, 106862. doi: 10.1016/j.ymssp.2020.106862

Kim, W. Y., Seo, H. I., and Seo, D. H. (2021). Nine-Axis IMU-based Extended inertial odometry neural network. *Expert. Syst.* 178, 115075. doi: 10.1016/j.eswa.2021.115075

Kong, J., Fan, X., Jin, X., Lin, S., and Zuo, M., (2023). A variational bayesian inference-based En-Decoder framework for traffic flow prediction. *IEEE Trans. Intell. Transp. Syst.* doi: 10.1109/TITS.2023.3276216

Kong, J.-L., Fan, X.-M., Jin, X.-B., Su, T.-L., Bai, Y.-T., Ma, H.-J., et al. (2023). BMAE-Net: a data-driven weather prediction network for smart agriculture. *Agronomy.* 13, 625. doi: 10.3390/AGRONOMY13030625

Li, D., Wu, Y., and Zhao, J. (2020). Novel hybrid algorithm of improved CKF and GRU for GPS/INS. *IEEE Access.* 8, 202836–202847. doi: 10.1109/ACCESS.2020.3035653

Liu, Q., Chen, Q. W., and Zhang, Z. C. (2017). Study on the fireman integrated positioning. *Techniq. Automat. Applicat.* 36, 34–37.

Liu, W., Gu, M., Mou, M., Hu, Y., and Wang, S. (2021). A distributed GNSS/INS integrated navigation system in a weak signal environment. *Meas. Sci. Technol.* 32, 115108. doi: 10.1088/1361-6501/ac07da

Liu, Y. H., Luo, Q. S., and Zhou, Y. M. (2022). Deep learning-enabled fusion to bridge GPS outages for INS/GPS integrated navigation. *IEEE Sens. J.* 22, 8974–8985. doi: 10.1109/JSEN.2022.3155166

Markus, W., Nils, G., and Michael, M. (2008). Systematic accuracy and precision analysis of video motion capturing systems-exemplified on the Vicon-460 system. *JBC.* 41, 2776–2780. doi: 10.1016/j.jbiomech.2008.06.024

Masiero, A., Guarnieri, A., Pirotti, F., and Antonio, V. (2014). A particle filter for smartphone-based indoor pedestrian navigation. *Micromachines-Basel* 5, 1012–1033. doi: 10.3390/mi5041012

Niu, H., and Lian, B. W. (2017). An integrated positioning method for GPS+PDR based on improved UKF filtering. *Bull. Surv Map* 2017:5–9. doi: 10.13474/j.cnki.11-2246.2017.0213

Pimpalkhute, V. A., Page, R., Kothari, A., Bhurchandic, K., and Kambled, V. (2021). Digital image noise estimation using DWT coefficients. *IEEE Trans.* 30, 1962–1972. doi: 10.1109/TIP.2021.3049961

Poulose, A., and Han, D. S. (2019). Hybrid indoor localization using IMU sensors and smartphone camera. *IEEE Sens. J.* 19, 5084. doi: 10.3390/s19235084

Skog, I., Nilsson, J. O., Zachariah, D., and Handel, P. (2013). Fusing the information from two navigation systems using an upper bound on their maximum spatial separation. *Int. Conferen. Indoor Position. Indoor Navigat.* 12, 862. doi: 10.1109/IPIN.2012.6418862

Soni, R., and Trapasiya, S. (2021). A survey of step length estimation models based on inertial sensors for indoor navigation systems. *Int. J. Commun. Syst.* 35, 5053. doi: 10.1002/dac.5053

Vidakovic, B., and Lozoya, C. B. (1998). On time-dependent wavelet denoising. *IEEE T Sign. Proces.* 46, 2549–2554. doi: 10.1109/78.709544

Wang, K. L. (2018). *The method research of indoor and outdoor pedestrians seamless navigation based on GNSS and IMU*. master degree, Nanchang University, Nanchang.

Wang, Q., Farahat, A., Gupta, C., and Zheng, S. (2021). Deep time series models for scarce data. *Neurocomputing.* 12, 132. doi: 10.1016/j.neucom.2020.12.132

Wang, X. (2019). *Research on key technology of image and wireless signal fusion location based on deep learning.* master degree, Beijing University of Posts and Telecommunications, Beijing.

Wang, Y., Liu, J., Li, R., and Suo, X. (2022b). Medium and long-term precipitation prediction using wavelet decomposition-prediction-reconstruction model. *Water Resour. Manag.* 37, 1473–1483. doi: 10.1007/s11269-022-03063-x

Wang, Y., Wang, H., Li, Q., Xiao, Y., and Ban, X. (2022a). Passive sonar target tracking based on deep learning. *J. Mar. Sci. Eng* 10, 181. doi: 10.3390/jmse10020181

Wu, X. Q., Lu, X. S., Wang, S. L., Wang, M. H., and Chai, D. S. (2020). A GNSS/INS integrated navigation algorithm based on modified adaptive kalman filter. *IJSTE* 20, 913–917. doi: 10.3969/j.issn.1671-1815.2020.03.007

Yang, Y. X. (2019). Deep learning-based cellular signal indoor localization algorithm. *J. CAEIT* 14, 943–947. doi: 10.3969/j.issn.1673-5692.2019.09.008

Zhang, J. M., Xiu, C. D., Yang, W., and Yang, D. (2018). Adaptive threshold zero-velocity update algorithm under multi-movement patterns. *J. Univ. Aeronaut. Astronaut.* 44, 636–644. doi: 10.13700/j.bh.1001-5965.2017.0148

Zhang, S. (2021). Challenges in knn classification. *T-KDE.* 99, 1–1. doi: 10.1109/TKDE.2021.3049250

Zhao, S., Zhou, Y., and Huang, T. C. (2022). A novel method for AI-assisted INS/GNSS navigation system based on CNN-GRU and CKF during GNSS Outage. *Remote Sensing* 14, 18. doi: 10.3390/rs14184494

Zheng, L., Zhouy, W., Tang, W., Zheng, X., Peng, A., and Zheng, H. (2016). A 3D indoor positioning system based on low-cost MEMS sensors. *Simul. Model Pract. TH.* 65, 45–56. doi: 10.1016/j.simpat.2016.01.003

Zhu, X. Y., Tao, T. Y., and Jiang, D. Z. (2018). Indoor positioning for firefighters based on GPS/MEMS inertial sensors. *J. H. Univ. Techno.* 41, 949–955. doi: 10.3969/j.issn.1003-5060.2018.07.016

frontiers | Frontiers in Neurorobotics

TYPE Original Research
PUBLISHED 14 July 2023
DOI 10.3389/fnbot.2023.1219170

Check for updates

OPEN ACCESS

EDITED BY
Long Jin,
Lanzhou University, China

REVIEWED BY
Ruoxi Qin,
Henan Key Laboratory of Imaging and Intelligent Processing, China
Hao Xu,
Anhui University of Technology, China

*CORRESPONDENCE
Yong Liu
✉ liuy1602@njust.edu.cn

RECEIVED 08 May 2023
ACCEPTED 26 June 2023
PUBLISHED 14 July 2023

CITATION
Wang K, Liu Y and Huang C (2023) Active fault-tolerant anti-input saturation control of a cross-domain robot based on a human decision search algorithm and RBFNN. *Front. Neurorobot.* 17:1219170.
doi: 10.3389/fnbot.2023.1219170

Active fault-tolerant anti-input saturation control of a cross-domain robot based on a human decision search algorithm and RBFNN

Ke Wang, Yong Liu* and Chengwei Huang

School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China

This article presents a cross-domain robot (CDR) that experiences drive efficiency degradation when operating on water surfaces, similar to drive faults. Moreover, the CDR mathematical model has uncertain parameters and non-negligible water resistance. To solve these problems, a radial basis function neural network (RBFNN)-based active fault-tolerant control (AFTC) algorithm is proposed for the robot both on land and water surfaces. The proposed algorithm consists of a fast non-singular terminal sliding mode controller (NTSMC) and an RBFNN. The RBFNN is used to estimate the impact of drive faults, water resistance, and model parameter uncertainty on the robot and the output value compensates the controller. Additionally, an anti-input saturation control algorithm is designed to prevent driver saturation. To optimize the controller parameters, a human decision search algorithm (HDSA) is proposed, which mimics the decision-making process of a crowd. Simulation results demonstrate the effectiveness of the proposed control methods.

KEYWORDS

cross-domain robot (CDR), radial basis function neural network (RBFNN), active fault-tolerant control (AFTC), anti-input saturation, human decision search algorithm (HDSA)

1. Introduction

In recent years, there has been a growing interest in multi-environment robots as single-environment robots are no longer sufficient to meet various practical needs (Cohen and Zarrouk, 2020). Researchers have proposed different designs to achieve this, such as bionic robots (Chen et al., 2021) and the legged amphibious robot (Xing et al., 2021). Furthermore, with the advancements in rotorcraft unmanned aerial vehicle (UAV) technology, researchers have started exploring the potential of integrating rotorcraft UAVs with wheeled mobile robots (WMRs) (Wang et al., 2019a). To enhance the capabilities of robots, cross-domain robots (CDRs) have been designed, which are capable of operating in multiple environments, including water, land, and air (Guo et al., 2019; Zhong et al., 2021). The robot presented in this paper is a CDR that combines a quadrotor UAV with a WMR equipped with webbed plates. These webbed plates on the wheels enable the robot to generate power at the water surface through their interaction with the water (Wang et al., 2022a,b).

The CDR presented in this study employs the same drive motors for ground and water surface operations. Assuming proper functionality during ground motion, a driver fault is considered to have occurred during the robot's operation on the water surface.

Fault-tolerant controls (FTCs) are control algorithms that effectively deal with system faults (Najafi et al., 2022; Nan et al., 2022). Sliding-mode controllers (SMCs) are commonly employed in passive fault-tolerant algorithms due to their robustness in maintaining control performance when the maximum system fault is known. However, the use of non-singular terminal sliding mode control (NTSMC) and SMC results in jitter problems, and this robust control approach is considered too conservative (Ali et al., 2020; Hou and Ding, 2021; Guo et al., 2022). To address these issues, FTCs frequently employ adaptive sliding mode control (Wu et al., 2020) and integral sliding mode control (Yu et al., 2022). Additionally, observers are commonly used to detect drive faults. In Wang F. et al. (2022), a disturbance observer (DO) is used to quickly compensate and correct unknown actuator faults of unmanned surface vehicles (USVs). In the context of autonomous underwater vehicles (AUVs), a sliding mode observer-based fault-tolerant control algorithm has been proposed in the literature (Liu et al., 2018). However, the design of higher-order observers requires complex mathematical proofs and the adjustment of many parameters. Neural networks (NNs) are often used to estimate system model parameters and uncertainty terms due to their ability to approximate arbitrary non-linear functions. In Zhang et al. (2022), NNs are used to rectify the model parameters of a USV, and an NN-based adaptive observer is developed to estimate errors caused by drive faults. As demonstrated in Gao et al. (2022), NNs can directly estimate system faults by approximating the uncertainty terms in the system. Event-triggered fault-tolerant control is a type of AFTC algorithm that has the potential to reduce system hardware requirements. However, it requires the development of trigger thresholds and corresponding fault control algorithms, which increase the difficulty and complexity of controller design (Huang et al., 2019; Wu et al., 2021; Zhang et al., 2021). Another important consideration in the FTC algorithm is the control of input saturation. One efficient approach for solving this issue is to introduce virtual states in the controller. These virtual states regulate the input error of the controller, thereby suppressing control input saturation (Wang and Deng, 2019). Additionally, designing adaptive laws is an effective way to address control input saturation. In this approach, the adaptive control input decreases as the actual control input approaches the maximum physical constraint (Shen et al., 2018).

The controller design presented above does not involve any optimization of the controller parameters. To address this limitation, reinforcement learning techniques have been developed to optimize control parameters. In Gheisarnejad and Khooban (2020), a reinforcement learning algorithm is employed to optimize the PID controller parameters. Another study (Zhao et al., 2020) trains the optimal trajectory following controller using deep reinforcement learning. However, reinforcement learning algorithms typically require a significant amount of data and multiple iterations to achieve optimal results. Swarm intelligence (SI) optimization algorithms are a promising approach in practical applications, including data classification, path planning, and controller optimization (Xue and Shen, 2020, 2022). Among the various SI optimization algorithms, particle swarm optimization (PSO) is a classical algorithm known for fast convergence and few parameters (Song and Gu, 2004). However, traditional PSO algorithms tend to fall into local optima. Ant colony optimization (ACO) is another common SI optimization algorithm. ACO can jump out of local optima but has slower convergence (Dorigo et al., 1996). In addition, the gray wolf optimizer (GWO) simulates the predation process of wolves (Mirjalili et al., 2014) and the Harris hawk optimizer (HHO) simulates the predation process of hawks (Heidari et al., 2019). These algorithms have shown improvements in convergence speed and accuracy compared with other animal predation simulation algorithms. Other popular SI optimization algorithms include the firefly algorithm (Fister et al., 2013) and the sine/cosine search algorithm (Mirjalili, 2016). Each SI optimization algorithm has its own strengths and weaknesses and no single algorithm can effectively handle all optimization problems. The goal is to achieve satisfactory results in terms of convergence speed, accuracy, and robustness for a specific optimization problem.

Based on the previous discussion, an AFTC is proposed for the CDR on the ground and on the water surface. This control algorithm consists of three main parts:

a. To enhance the robustness of the robot control system, a fast NTSMC is designed based on the concept of passive FTC. Compared with traditional NTSMC and SMC, the proposed NTSMC has reduced control input chatter. Additionally, to reduce controller conservatism, an RBFNN is designed to detect and compensate for drive faults. The adaptive weight control law of the RBFNN is based on the Lyapunov function.
b. To prevent drive saturation, an anti-input saturation control algorithm based on the hyperbolic tangent (tanh) function is employed. An adaptive rate is designed to prevent singularities in this algorithm. This method does not require complex mathematical proofs and requires fewer tuning parameters.
c. A new SI optimization algorithm named HDSA is proposed for the optimization of the weight update rate parameter of RBFNNs. The proposed algorithm is compared with other SI optimization algorithms, and the test results demonstrate its faster convergence rate and higher accuracy.

2. Related work and mathematical models

2.1. HDSA's related work

To demonstrate the advantages of the proposed HDSA optimization algorithm, the results of the HDSA tests are shown in this section. The theory of HDSA is discussed in detail in the section entitled "RBFNN-Based Active Fault-Tolerant Control Algorithm". The effectiveness of the proposed optimization algorithm was evaluated by comparing the test results of HDSA with other popular optimization algorithms, such as particle swarm optimization (PSO) (Song and Gu, 2004), the sine/cosine algorithm (SCA) (Mirjalili, 2016), the gray wolf optimizer (GWO) (Mirjalili et al., 2014), the firefly algorithm (FA) (Fister et al., 2013), and the Harris hawk optimizer (HHO) (Heidari et al., 2019). Twenty standard test functions were used for evaluation, which are presented in Tables 5–7 (included in the Simulation Results section).

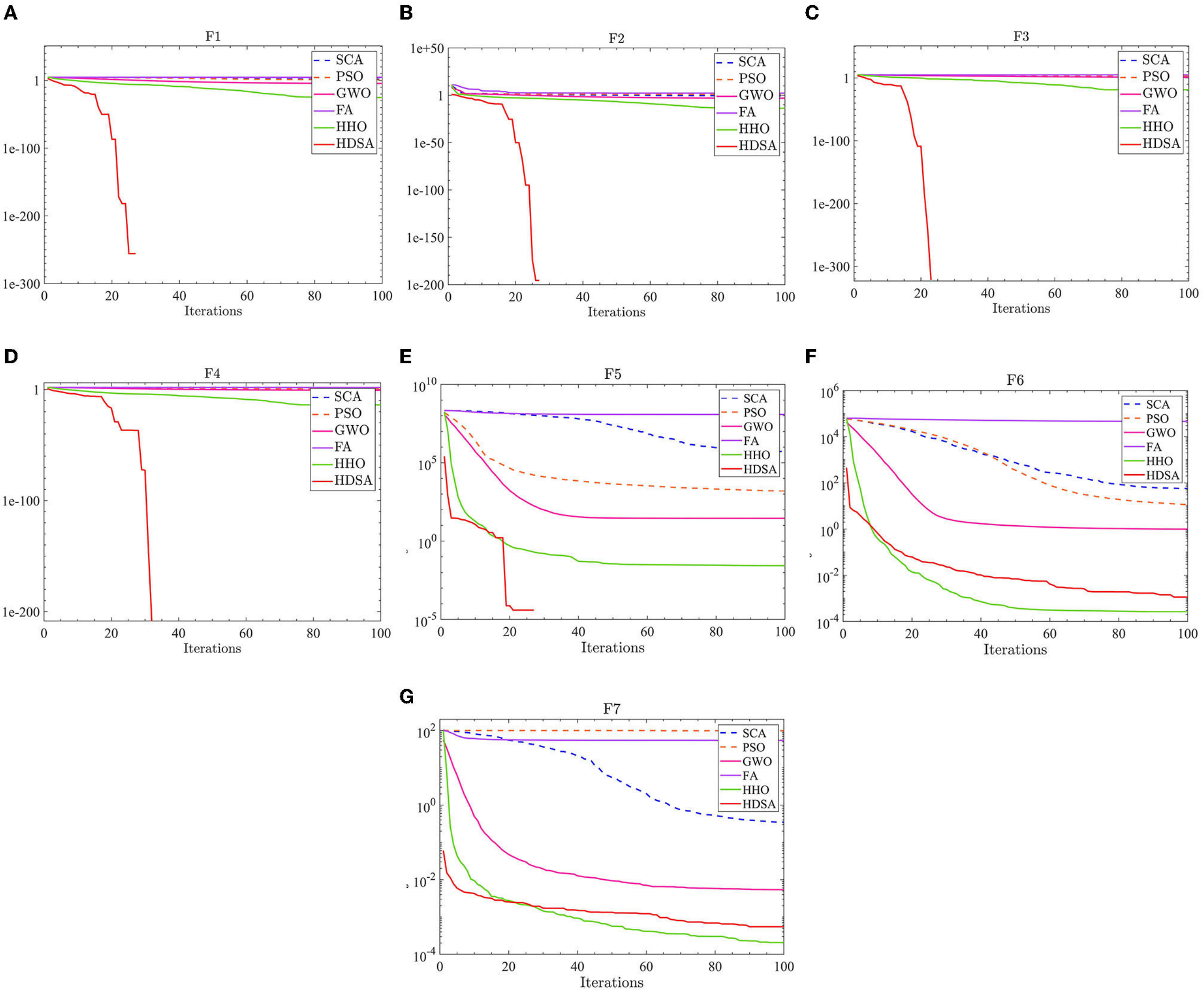


FIGURE 1
Single peak function test results. **(A–G)** represent the test results of the six algorithms in functions F1 to F7.

The number of populations was $pop = 100$ and the maximum number of iterations was $M = 100$. The average fitness over 30 independent runs was considered as the optimization result. The convergence characteristics of the six algorithms in the single-peak function test are depicted in Figure 1, while Figure 2 illustrates the convergence characteristics in the multi-peak function test. Furthermore, Figure 3 demonstrates the convergence characteristics of the six algorithms on fixed-dimensional multi-peak functions. The test results of the six algorithms, based on 30 independent runs, are summarized in Tables 1, 2. In Tables 1, 2, purple indicates the optimal value of the test functions, pink indicates the mean value of the test functions, and white indicates the mean squared deviation of the test functions.

The results of the single-peak functions F1–F7 test results are presented in Tables 1, 2. In these tests, the mean and optimal values obtained by HDSA in F1–F5 are both 0, indicating that HDSA achieves the highest accuracy among the six algorithms. Although the accuracy of HDSA is slightly inferior to HHO in the F6–F7 test functions, it still outshines SCA, PSO, GWO, and FA. HDSA has a standard deviation of 0 in tests F1–F5, suggesting that HDSA is the most stable algorithm. Although its stability is slightly lower than HHO in tests F6–F7, it still outperforms the other four methods. Convergence speed is depicted in Figure 2. HDSA has a significantly faster convergence speed compared with the other five algorithms, but its convergence accuracy in the F6–F7 tests is lower than that of HHO.

The test results for the multi-peak functions F8–F13 are presented in Tables 1, 2. In the tests from F9 to F13, HDSA exhibits significantly better stability and convergence accuracy compared with the other five algorithms. It achieves higher accuracy and the smallest standard deviation. As depicted in Figure 3, except for the F8 test function, HDSA showcases the fastest convergence speed and highest convergence accuracy among the algorithms.

The results of the fixed dimensional multi-peak functions F14–F20 test results are shown in Tables 1, 2. In the F14 test, SCA has the best optimal and average accuracy, while HDSA exhibits slightly lower average accuracy and stability compared with SCA, PSO, and HHO. However, HDSA still manages to find the optimal

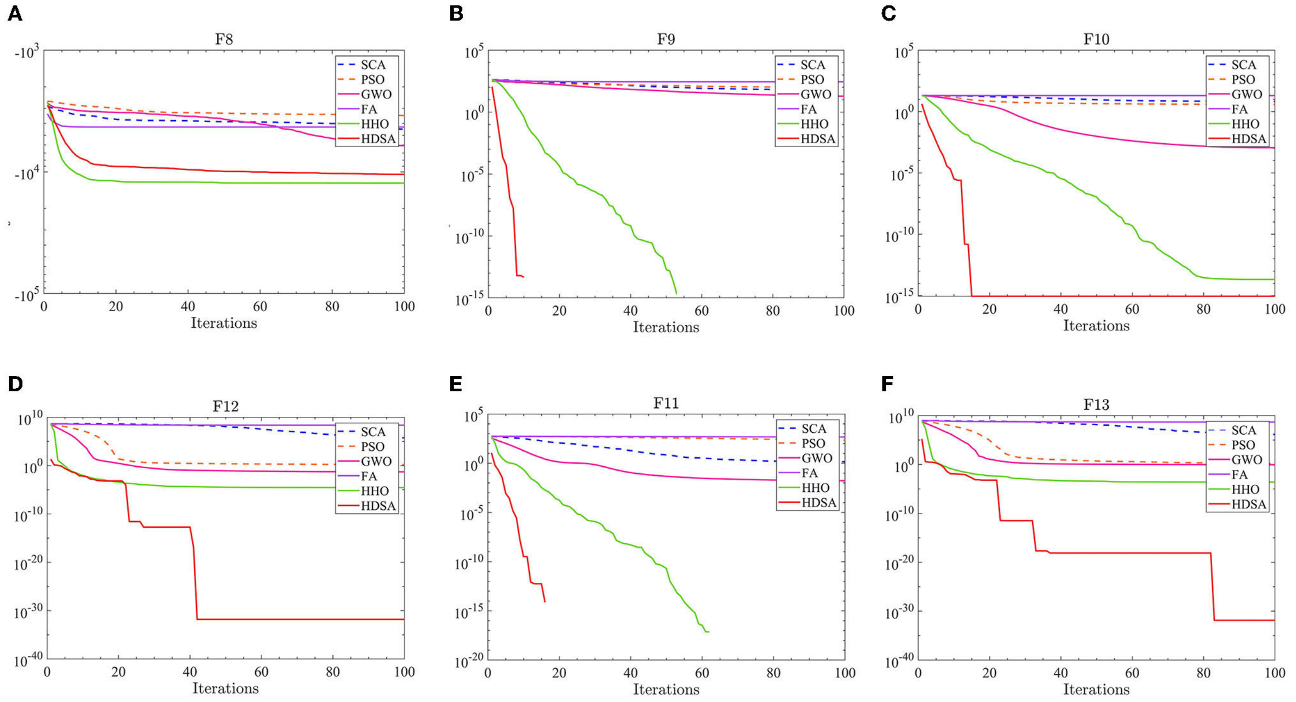


FIGURE 2
Multi-peak function test results. **(A–F)** represent the test results of the six algorithms in functions F8 to F13.

solution in 30 runs. In the F15–F18 test results, HDSA, SCA, GWO, and HHO perform closely, with good stability and accuracy. In the F19–F20 tests, HDSA outperforms the other five algorithms significantly in terms of accuracy and stability. As shown in Figure 3, HDSA exhibits the fastest convergence speed among the other test functions, except for F15, F17, and F18. In the F15 test, HDSA is only slightly slower than HHO, while in the F17 and F18 tests, HDSA converges slightly slower than FA.

2.2. Mathematical model of the CDR

Before discussing the mathematical model of the CDR, the following assumptions are made: **Assumption 1:** The center of gravity and the geometric center of the robot body coincide. **Assumption 2:** The motor output torque meets the actual performance requirements of the robot during ground and water motion. **Assumption 3:** The robot's vertical swing, horizontal rocking, and longitudinal rocking during its movement on the water surface are ignored. **Assumption 4:** The motion of the robot on the ground is purely rolling, without any sliding motion.

The CDR designed in this study can be seen as a combination of a quadrotor UAV and a WMR. Figure 4A shows the robot moves on the ground. Figure 4B shows the robot moves on the water surface by webbed plates. Figure 4C shows the robot moves on the water surface by proplers. The robot moves in the air in a similar way to the quadrotor UAV as shown in Figures 4D, E. Figure 4F shows the structure of the robot, where webbed plates are mounted on the wheels. These webbed plates generate traction and rotational torque on the water surface by interacting with the water. However, as this paper focuses primarily on the FTC algorithm of the robot on the ground and on the water surface, the discussion does not explore the robot's aerial motion in detail.

The robot in the inertial frame and in the body frame is shown in Figure 5.

In Figure 5, d is the distance from the geometric center of the robot O_b to the mass center of the robot. b is the axis radius and r is the wheel radius. ω_l, ω_r are the angular velocities of the left and right wheels. ψ is the angle between the robot body coordinate system b and the inertial coordinate system A, and ψ is the yaw angle of the robot. The kinematic model of the robot on the ground and water surface can be represented as (Liu et al., 2020):

$$\dot{q} = R\eta \tag{1}$$

where $q = \left[\begin{array}{ccc} x & y & \psi \end{array}\right]$ represents the position and orientation of the robot in the inertial frame, while $\eta = \left[\begin{array}{ccc} u & v & r \end{array}\right]$ is used to denote the longitudinal velocity, lateral velocity, and yaw angular velocity in the body frame. The coordinate conversion matrix is denoted by R, where $R=\left[\begin{array}{ccc} \cos\psi & \sin\psi & 0 \\ -\sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{array}\right]$. The dynamics model of the robot's motion on the ground can be expressed as

$$M(q)\ddot{q} + C_m(q,\dot{q})q + F(\dot{q}) + \tau_d = B(q)\tau \tag{2}$$

The matrices M are symmetric positive definite inertia matrices, while C_m represents the centripetal and Coriolis matrix. The term $F(\dot{q})$ denotes mechanical friction, while τ_d is used to represent external disturbances. The input transformation matrices are

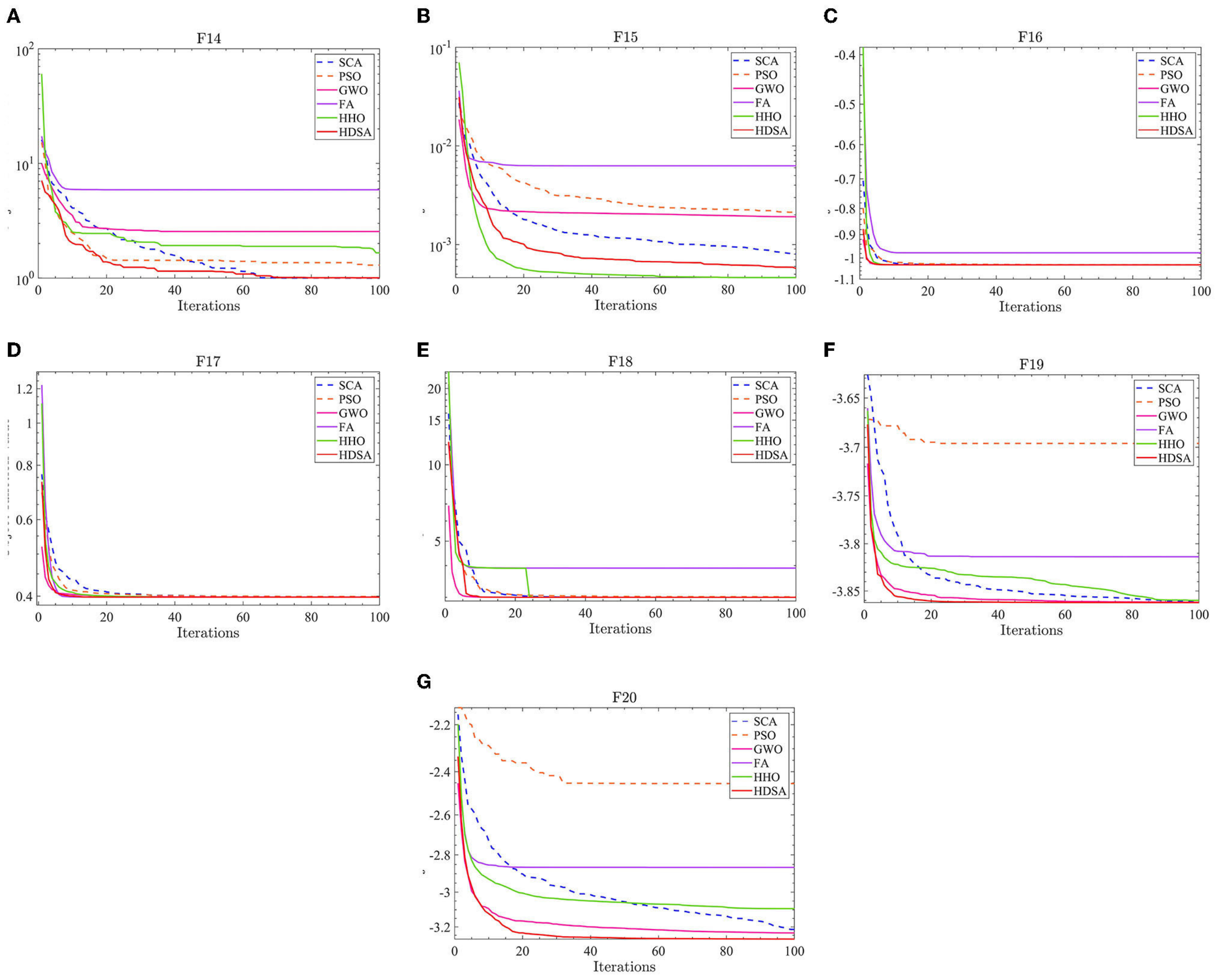


FIGURE 3
Fixed dimensional multi-peak function results. **(A–G)** represent the test results of the six algorithms in functions F14 to F20.

denoted as $B(q)$. Furthermore, the robot drive motors in the left and right wheel output torque are represented by $\tau = \begin{bmatrix} \tau_l & \tau_r \end{bmatrix}^T$.

$$M(q) = \begin{bmatrix} m & 0 & md\sin\psi \\ 0 & m & -md\cos\psi \\ md\sin\psi & -md\cos\psi & I \end{bmatrix},$$

$$B(q) = \frac{1}{r}\begin{bmatrix} \cos\psi & \cos\psi \\ \sin\psi & \sin\psi \\ L & -L \end{bmatrix},$$

$$C_m(q,\dot{q}) = \begin{bmatrix} md\dot{\psi}^2\cos\psi & md\dot{\psi}^2\sin\psi & 0 \end{bmatrix}^T$$

The mass of the robot is represented by m. The I is a scalar quantity and represents the rotational inertia of the robot as it rotates in the X-Y plane. The angular velocity of the robot is assumed to vary smoothly, so that $\dot{\psi} \approx 0$. According to **assumption 1**, the Coriolis matrix can be assumed to be negligible, resulting in $C_{\mathrm{m}} \approx 0$. According to **assumption 1**, $d = 0$, so the matrix $M(q) = diag\begin{bmatrix} m & m & I \end{bmatrix}$. Based on these assumptions, the dynamics model of the robot on the ground can be rewritten as follows:

$$\bar{M}\ddot{q} + \bar{C}q + +\bar{F}(\dot{q}) + \bar{\tau}_d = \bar{B}\tau \tag{3}$$

where $\bar{C} = R^{-1}C_m\dot{R}$, $\bar{M} = R^{-1}MR$, $\bar{B} = R^{-1}B$. $\bar{F}(\dot{q}) = \begin{bmatrix} f_u & f_v & f_r \end{bmatrix}^T$ is the mechanical friction and $\bar{\tau}_d = \begin{bmatrix} d_u & d_v & d_r \end{bmatrix}^T$ is the external disturbance. Rewriting 3 into algebraic form can be expressed as:

$$\begin{cases} \dot{u} = \left(F_u - f_u - d_u\right)/m + v\omega \\ \dot{v} = -u\omega - \left(f_v + d_v\right)/m \\ \dot{r} = \left(T_r - f_r - d_r\right)/I \end{cases} \tag{4}$$

The traction force is represented by F_u, while T_r represents the torque. To model the dynamics of the robot on the water surface, we can refer to the USV dynamics model (Chen et al., 2019), which can be expressed as follows

$$M_w(q)\dot{\eta} + C_w\left(q,\eta\right) + D_w\left(\eta\right)\eta + F_w(\eta) + \tau_{dw} = \tau_w \tag{5}$$

TABLE 1 Test results of HDSA SCA and PSO algorithms run independently 30 times.

F	HDSA			SCA			PSO		
	Best	Ave	Std	Best	Ave	Std	Best	Ave	Std
F1	0	0	0	0.043621199	55.2091773	36.42876978	4.456913956	8.976645307	46.32264983
F2	0	0	0	0.010156561	0.288456338	0.399117642	6.86827201	10.3267055	10.28163277
F3	0	0	0	9.53E+03	2.23E+04	5.86E+03	2.70E+02	7.57E+02	2.15E+04
F4	0	0	0	37.35838437	59.20677056	8.010506582	1.92884739	3.830868295	55.38478266
F5	0	0	0	37.92815456	6.61E+05	8.02E+05	5.46E+02	1.57E+03	6.59E+05
F6	3.14E−05	0.001383925	0.001405319	4.67425676	1.15E+02	1.15E+02	6.428054849	10.03538963	57.17359472
F7	5.90E−05	5.11E−04	4.07E−04	0.024658139	0.341143612	0.269446788	45.07882375	96.08387995	98.09394285
F8	−1.26E+04	−1.07E+04	1.97E+03	−4.81E+03	−4.37E+03	2.21E+02	−4.12E+03	−3.49E+03	9.36E+02
F9	0	0	0	0.860127299	78.49328591	70.22545167	30.55768733	94.82695388	34.53524407
F10	8.88E−16	8.88E−16	0	0.187682845	10.65598272	8.935131413	2.867362804	3.884021036	6.796254632
F11	0	0	0	0.513962142	1.962336683	2.819722656	1.72E+02	2.25E+02	2.24E+02
F12	1.57E−32	1.57E−32	5.47E−48	1.043279428	3.39E+05	8.56E+05	0.650894524	1.738034681	3.39E+05
F13	1.35E−32	1.84E−23	9.89E−23	10.16366581	2.10E+06	2.56E+06	0.62640203	1.796606655	2.10E+06
F14	0.998003838	4.801561855	4.696216357	0.998003841	0.998323781	9.29E−04	0.998003838	1.163740602	0.405679435
F15	3.08E−04	4.82E−04	2.60E−04	4.25E−04	8.05E−04	1.88E−04	5.35E−04	0.003654847	0.007160687
F16	−1.031628435	−1.03162038	8.23E−06	−1.031628443	−1.031626913	1.82E−06	−1.031615014	−1.031069614	6.31E−04
F17	0.397888187	0.397903308	1.88E−05	0.397889317	0.397918592	3.12E−05	0.397935785	0.399283994	0.001760397
F18	3.000000032	3.000013391	1.62E−07	3.000000177	3.000055929	8.99E−05	3.000002051	3.007764558	0.014062845
F19	−3.862751312	−3.862443601	2.78E−04	−3.86268097	−3.861957813	0.00102005	−3.849759489	−3.653030339	0.272064362
F20	−3.320685667	−3.277232199	0.056398698	−3.314075954	−3.22298511	0.922404083	−2.942883457	−2.387193028	0.041529991

M_w is the inertia matrix. The traction force and torque of the robot at the water surface are $\tau_w = \left[F_u \ 0 \ T_r \right]^T$. $\tau_{dw} = \left[d_{uw} \ d_{vw} \ d_{rw} \right]^T$ is the lumped disturbance and $F_w(\eta)=\left[f_{uw} \ f_{vw} \ f_{rw} \right]$ is the water resistance.

$$M_w = \begin{bmatrix} m_{11} & 0 & 0 \\ 0 & m_{22} & m_{23} \\ 0 & m_{32} & m_{33} \end{bmatrix},$$

$$C_w\left(q,\eta\right) = \begin{bmatrix} 0 & 0 & C_{13}\left(\eta\right) \\ 0 & 0 & C_{23}\left(\eta\right) \\ -C_{13}\left(\eta\right) & -C_{23}\left(\eta\right) & 0 \end{bmatrix},$$

$$D_w\left(\eta\right) = \begin{bmatrix} d_{11} & 0 & 0 \\ 0 & d_{22} & d_{23} \\ 0 & d_{32} & d_{33} \end{bmatrix}.$$

The disturbances are represented by τ_{dw}. On the other hand, $D_w\left(\eta\right)$ represents the water resistance. The Coriolis force matrix can also be neglected according to **Assumption 1** and **Assumption 3**, so $C_w\left(q,\eta\right) \approx 0$. The elements of the non-diagonal matrix in matrix $D_w\left(\eta\right)$ and matrix M_w are small and can be neglected. This model simplification approach is also more common (Liao et al., 2016; Wang et al., 2019b; Deng et al., 2020), where $m_{11} = m - X_{\dot{u}}$, $m_{22} = m - Y_{\dot{v}}$, and $m_{33} = I_z - N_{\dot{r}}$ are the inertia parameters of the three axes and $X_{\dot{u}}$, $Y_{\dot{v}}$, and $N_{\dot{r}}$ are the additional inertia parameters due to the wet water of the robot shell and the viscosity of the water. The dynamics model of the robot on the water surface can be expressed as:

$$\begin{cases} \dot{u} = \frac{m_{22}}{m_{11}} v\omega - \frac{X_u}{m_{11}} u - \frac{X_{|u|u}}{m_{11}} \left|u\right| u + \frac{F_u}{m_{11}} + \frac{d_u}{m_{11}} \\ \dot{v} = -\frac{m_{11}}{m_{22}} u\omega - \frac{Y_u}{m_{22}} v - \frac{Y_{|v|v}}{m_{22}} \left|v\right| v + \frac{d_v}{m_{11}} \\ \dot{\omega} = \frac{m_{11}-m_{22}}{m_{33}} u\upsilon - \frac{N_\omega}{m_{33}} \omega - \frac{N_{|\omega|\omega}}{m_{33}} \left|\omega\right| \omega + \frac{T_r}{m_{33}} + \frac{d_r}{m_{33}} \end{cases} \tag{6}$$

X_u, $X_{|u|u}$, Y_u, $Y_{|v|v}$, and N_ω, $N_{|\omega|\omega}$ are the resistance coefficients. The resistance of the robot moving on the water surface can be approximated as a quadratic function of the velocity and angular velocity.

The mathematical model should be rewritten into a form that better suits the needs of the subsequent controller design. The dynamics model of the robot's motion on the ground is rewritten according to 4 as

$$\begin{cases} \dot{u} = F_u/m - \underbrace{\left(f_u + d_u\right)/m + v\omega}_{d_{ug}} \\ \dot{r} = T_r/I - \underbrace{\left(f_r + d_r\right)/I}_{d_{rg}} \end{cases} \tag{7}$$

TABLE 2 Test results of GWO FA and HHO algorithms run independently 30 times.

F	GWO			FA			HHO		
	Best	Ave	Std	Best	Ave	Std	Best	Ave	Std
F1	2.69E−06	2.59E−05	1.52E−05	2.29E+04	4.77E+04	9.85E+03	1.08E−33	1.95E−26	7.53E−26
F2	4.61E−04	8.55E−04	2.86E−04	53.06138425	1.06E+02	18.47075013	1.08E−17	2.26E−14	6.46E−14
F3	2.333236967	16.15496494	16.17029995	3.37E+04	6.69E+04	1.69E+04	6.30E−32	4.03E−18	2.17E−17
F4	0.076161242	0.190101216	0.067944463	46.02316618	63.05292108	7.574286394	1.03E−17	4.60E−14	1.18E−13
F5	26.18035457	28.03756308	0.956956464	5.84E+07	1.29E+08	3.79E+07	1.17E−04	0.043168201	0.061375953
F6	3.62E−04	0.996486127	0.498192351	3.06E+04	4.64E+04	7.31E+03	2.08E−06	2.68E−04	3.23E−04
F7	0.001822731	0.004742428	0.001594265	10.11397979	48.34775676	17.25308781	9.10E−06	1.82E−04	1.50E−04
F8	−8.30E+03	−6.30E+03	1.05E+03	−5.73E+03	−4.35E+03	6.34E+02	−1.26E+04	−1.25E+04	2.43E+02
F9	10.61773399	21.78763545	6.854472334	2.15E+02	0.860127299	34.00946589	0	0	0
F10	6.79E−04	0.001208914	4.59E−04	19.41517193	19.96298677	0.1314286	8.88E−16	1.33E−14	2.00E−14
F11	2.16E−05	0.020113329	0.01797284	4.03E+02	4.93E+02	48.66049354	0	0	0
F12	0.01735841	2.091212922	0.039891925	5.55E+07	2.26E+08	1.13E+08	2.47E−07	2.25E−05	2.25E−05
F13	0.39714087	0.90669375	0.26593137	1.29E+08	4.74E+08	1.87E+08	5.84E−11	3.13E−04	4.99E−04
F14	0.998003838	2.149370759	1.977247758	0.998003838	9.85228046	7.376397236	0.998003838	1.592846754	1.007706592
F15	3.33E−04	0.002524088	0.005948479	5.95E−04	0.009720032	0.008406408	3.09E−04	4.19E−04	2.61E−04
F16	−1.031628453	−1.031628406	8.86E−04	−1.031621754	−1.030900759	0.002366781	−1.031628453	−1.031628451	1.05E−08
F17	0.397887459	0.397888965	1.47E−06	0.397894813	0.398122914	3.37E−04	0.397887358	0.397893418	2.30E−05
F18	3.000000021	3.000091041	9.87E−05	3.000120892	3.027874998	0.065760186	3	3.000000968	4.43E−06
F19	−3.86278078	−3.861772215	0.00183244	−3.861890169	−3.830959144	0.086620017	−3.862769505	−3.861362289	0.001672668
F20	−3.321992055	−3.265460239	0.071106357	−3.201236207	−2.894366935	0.195231925	−3.263585483	−3.123254299	0.085277304

Where d_{ug} is the lumped disturbance and $d_{ug} \leq \bar{d}_{ug}$, $\bar{d}_{ug}$ is the upper limit of the total disturbances. d_{rg} is the lumped disturbance and $d_{rg} \leq \bar{d}_{rg}$, $\bar{d}_{rg}$ is the upper limit of the total disturbances. The dynamics model of the robot on the water surface is

$$\begin{cases} \dot{u} = \frac{F_{uc}}{m} \underbrace{- \frac{\xi_u F_{uc}}{m_{11}} - F_{ua} - \frac{X_u}{m_{11}} u - \frac{X_{|u|u}}{m_{11}} |u| \, u + \Delta_F}_{-D_{uw}} \\ \quad + \underbrace{\frac{m_{22}}{m_{11}} v\omega + \frac{d_u}{m_{11}}}_{d_{uw}} \\ \dot{r} = \frac{T_{rc}}{I} \underbrace{- \frac{\xi_r T_{rc}}{m_{33}} - T_{ra} - \frac{N_\omega}{m_{33}} \omega - \frac{N_{|\omega|\omega}}{m_{33}} |\omega| \, \omega + \Delta_T}_{-D_{rw}} \\ \quad + \underbrace{\frac{m_{11} - m_{22}}{m_{33}} uv + \frac{d_r}{m_{33}}}_{drw} \end{cases} \tag{8}$$

where F_{uc} is the desired tractive force and $F_{uc} = F_u$ represents no force loss. $\xi_u \in \left[0 \ 1 \right)$ is the force loss parameter. Δ_F is the force disturbance due to mass change. d_{uw} is a lumped disturbance, $d_{uw} \leq \bar{d}_{uw}$. $\bar{d}_{uw}$ is the upper bound of d_{uw}. D_{uw} is the uncertainty term when the robot moves on the water surface due to changes in system parameters, water resistance, and driver faults. T_{rc} is the desired torque and $T_{rc} = T_r$ represents no force loss. $\xi_r \in \left[0 \ 1 \right)$ is the power loss parameter. Δ_T is the torque disturbance due to the change of inertia parameter. d_{rw} is a lumped disturbance, $d_{rw} \leq \bar{d}_{rw}$. $\bar{d}_{rw}$ is the upper bound of d_{rw}. D_{rw} is the uncertainty term due to changes in system parameters, water resistance, and driver faults during robot rotation on the water surface.

3. Active fault tolerance control algorithm and human decision search algorithm

3.1. RBFNN-based active fault-tolerant control algorithm

Both the yaw control and the linear velocity control of the robot are essentially single-input single-output (SISO) second-order non-linear affine systems. Without loss of generality, a second-order non-linear affine SISO system with drive faults can be expressed as:

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = f(x) + g(x) u_c + D \ + d \\ y = x_1 \end{cases} \tag{9}$$

u_c is unconstrained control input, u_a is the drive bias, ξ is the power loss parameter, $\xi \in \left[0 \ 1 \right)$, 0 represents no power loss, and 1 represents a complete loss of efficiency. $D = -g(x)\xi u_c + u_a$ is the uncertainty term due to the driver fault. The disturbance d has a

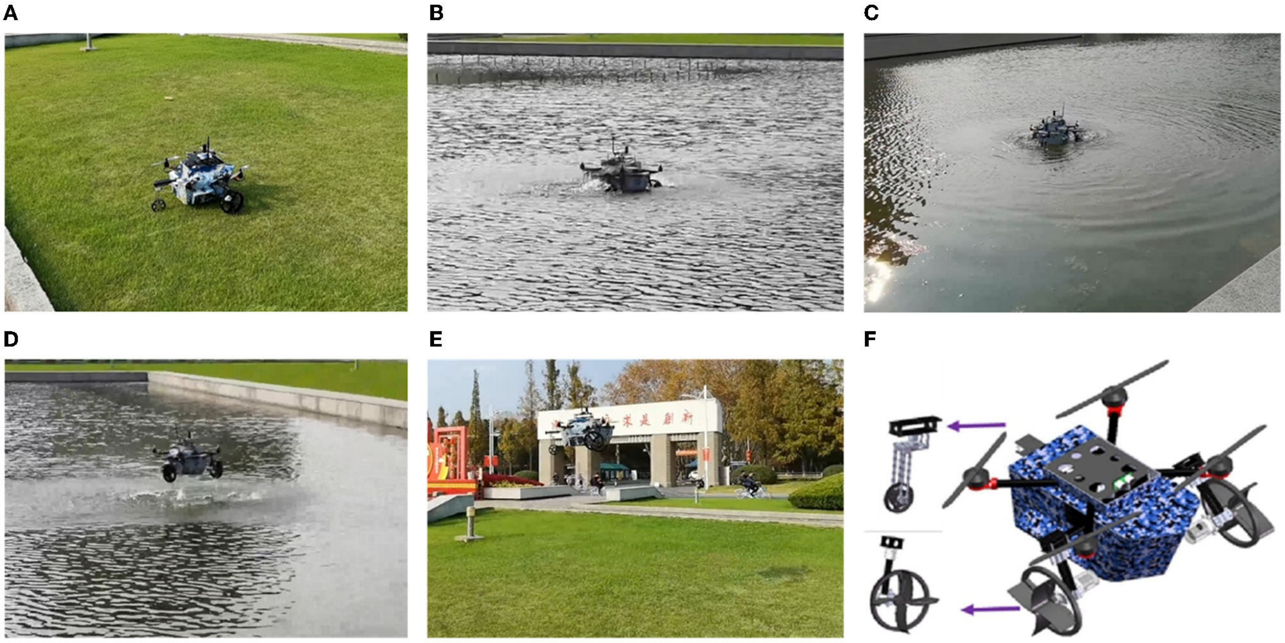


FIGURE 4
(A) The robot moves on the ground. **(B)** The robot moves on the water surface by webbed plates. **(C)** The robot moves on the water surface by propllers. **(D)** The robot takes off from water surface. **(E)** The robot flying in the air. **(F)** The structure of robot.

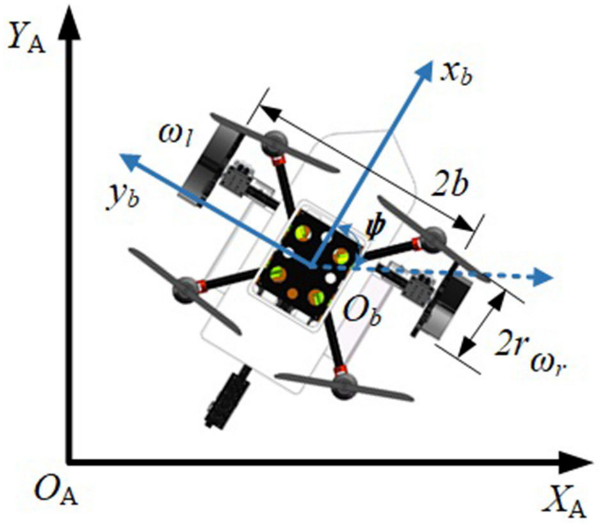


FIGURE 5
Robot in the inertial frame and the body frame.

well-defined upper limit and $|d| \le \bar{d}$. x_1, x_2 are system states. $f(x)$ is the system function and $g(x)$ is the input function. Owing to the physical constraints of the controlled object, the control input is subject to saturation:

$$u_{con} = \begin{cases} u_{\max}, & |u_c| > u_{\max} \\ u_c, & u_c \le u_{\max} \end{cases} \tag{10}$$

$u_{\max}$ is the physical constraint. To make the control input smoother, the cutoff function is usually replaced by a saturation function, such as *tanh*.

$$u_{con} = u_{\max} \tanh(u_f / u_{\max}) \tag{11}$$

where u_{con} is the constrained control input and u_f is a function of u_c. Thus, the control objective is to design the constrained control law u_{con} so that it satisfies the control requirements even in the presence of drive faults and external disturbances in the controlled object. The steps for designing an AFT controller are the following:

Step 1: Define the state error $e_1 = x_{1\mathrm{d}} - x_1$. Establish the Lyapunov function $V_1 = \frac{1}{2}e_1^2$. Taking the derivative of V_1 with respect to the time t gives

$$\dot{V}_1 = e_1\dot{e}_1 = e_1(\dot{x}_{1d} - x_2) \tag{12}$$

Define the virtual state $\alpha_x = k_1 e_1 + \dot{x}_{1d}$ as the desired input of the next step. If x_2 can follow α_x, $\dot{V}_1 = -k_1 e_1^2$. So, the next step of the control law must ensure that $\alpha_x - x_2 = 0$. α_x is the next desired state x_{2d}.

Step 2: Define the state error $e_2 = x_{2\mathrm{d}} - x_2$, and the fast NTSMC is designed as

$$S = e_2 + \alpha e_1 + \beta e_1^{\lambda} \tag{13}$$

where α and β are positive adjustable parameters and λ is a positive odd number. The sliding mode convergence law is

$$\dot{S} = -k_2 S - k_3 |S|^{\gamma_1} \operatorname{sgn}(S) \tag{14}$$

where k_1, k_2, and γ_1 are positive adjustable parameters. *sgn* is the symbolic function. The derivation of 13 yields:

$$\dot{S} = \dot{e}_2 + \alpha\dot{e}_1 + \lambda\beta {e_1}^{\lambda-1}\dot{e}_1 = -k_2S - k_3|S|^{\gamma_1}\operatorname{sgn}(S) \tag{15}$$

where

$$\begin{aligned}\dot{e}_2 &= \dot{x}_{2\mathrm{d}} - \dot{x}_2\\ &= \dot{\alpha}_x - f(x) - g(x)u_c - d - D\\ &= -k_2S - k_3|S|^{\gamma_1}\operatorname{sgn}(S)\end{aligned} \tag{16}$$

The controller law can be designed as follows:

$$u_c = \frac{1}{g(x)}\left(\dot{\alpha}_x - f(x) - D + k_2S + k_3|S|^{\gamma_1}\operatorname{sgn}(S) + \alpha\dot{e}_1 + \lambda\beta {e_1}^{\lambda-1}\dot{e}_1\right) \tag{17}$$

In 17, the uncertain term due to drive faults D is known. Establishing the Lyapunov function $V_2 = \frac{1}{2}S^2$, the derivative of V_2 yields

$$\begin{aligned}\dot{V}_2 &= S\dot{S}\\ &= S\left(\dot{e}_2 + \alpha\dot{e}_1 + \lambda\beta {e_1}^{\lambda-1}\dot{e}_1\right)\\ &= S\left(\dot{\alpha}_x - f(x) - g(x)u_c - d - D + \alpha\dot{e}_1 + \lambda_1\beta {e_1}^{\lambda_1-1}\dot{e}_1\right)\end{aligned} \tag{18}$$

Bringing 17 into 18 yields

$$\begin{aligned}\dot{V}_2 &= S\dot{S}\\ &= S\left(-d - k_2S - k_3|S|^{\gamma_1}\operatorname{sgn}(S)\right)\\ &= -k_2S^2 - k_3|S|^{\gamma_1+1} - Sd\\ &\le -k_2S^2 - k_3|S|^{\gamma_1+1} + |S|\,\bar{d}\\ &= -k_2S^2 - k_3|S|^{\gamma_1+1} + |S|\,\bar{d}\\ &= -k_2S^2 - |S|\left(k_3|S|^{\gamma_1} - \bar{d}\right)\end{aligned} \tag{19}$$

When $k_3 > \bar{d}/|S|^{\gamma_1}$, $k_3|S|^{\gamma_1} - \bar{d} = \varepsilon$, $\varepsilon > 0$, thus:

$$\dot{V}_2 \le -2k_2V_2 - \varepsilon\,|S| \le -2k_2V_2 - \sqrt{2}\varepsilon V_2^{1/2} < -\alpha_1 V_2^{1/2} - \beta_1 V_2 \tag{20}$$

where $\alpha_1 = 2k_2$, $0 < \beta_1 < \sqrt{2}\varepsilon$.

LEMMA 1 [44] (Jiang and Lin, 2020): Consider a smooth positive definite $V(x)$, $x \in R_n$. Suppose that real numbers $p_1 \in (0, 1)$, $\alpha > 0$, and $\beta > 0$ exist such that $V(x) < -\alpha V(x)^{p_1} - \beta V(x)$. Then, an area $U_0 \in R_n$ exists, such that any $V(x)$ starting from U_0 can reach $V(x) = 0$ in finite time T_v, which is expressed as $T_v \le \frac{1}{\beta(1-p_1)}\ln\left(\frac{V^{1-p_1}(x_0)+\alpha}{\alpha}\right)$.

According to **lemma 1**, V_2 can converge to 0 in finite time. In the above discussion, the uncertainty term D is assumed to be known, but the actual uncertain term D is unknown. As RBFNN can approximate arbitrary uncertain non-linear functions and does not depend on a mathematical model, it is more suitable for estimating stochastic uncertain terms. Therefore, optimal neural network weights w^* must exist such that $D{=}\varepsilon_0{+}{w^*}^T h$, ε_0 is the estimated residual and h is the neuron. $\tilde{w} = \hat{w} - w^*$, $\hat{w}$ is an estimate of w^* and w^* is a constant, so $\dot{\tilde{w}} = \dot{\hat{w}}$. Rewrite 9 as:

$$\begin{cases}\dot{x}_1 = x_2\\ \dot{x}_2 = f(x) + g(x)u_c + d + \varepsilon_0 + {w^*}^T h\\ y = x_1\end{cases} \tag{21}$$

Step 3: Establish the Lyapunov function V_3 as

$$V_3 = \frac{1}{2}S^2 + \frac{1}{2}tr(\tilde{w}^T\Gamma^{-1}\tilde{w}) \tag{22}$$

The derivation of formula 22 yields

$$\begin{aligned}V_3 &= S\dot{S} + \tilde{w}^T\Gamma^{-1}\dot{\hat{w}}\\ &= S\left(\dot{\alpha}_x - f(x) - g(x)u_c - d - \varepsilon_0 - {w^*}^T h + \alpha\dot{e}_1 + \lambda_1\beta e_1^{\lambda_1-1}\dot{e}_1\right)\\ &+ \tilde{w}^T\Gamma^{-1}\dot{\hat{w}}\end{aligned} \tag{23}$$

The control law is designed to

$$u_c = \frac{1}{g(x)}\left(\dot{\alpha}_x - f(x) - \hat{w}^T h + k_2S + k_3|S|^{\gamma_1}\operatorname{sgn}(S)\right) \tag{24}$$

Bringing formula 24 into 23 yields

$$\dot{V}_3 = -k_2S^2 - k_3|S|^{\gamma_1+1} - S\varepsilon_1 + \tilde{w}^T(Sh + \Gamma^{-1}\dot{\hat{w}}) \tag{25}$$

where $\varepsilon_1 = d + \varepsilon_0$, the upper limit of the estimation error of the neural network is $\bar{\varepsilon}_0$. $\bar{\varepsilon}_0 \ge \varepsilon_0$, $\bar{d} \ge d$, so that $\varepsilon_1 \le \bar{d} + \bar{\varepsilon}_0 = \bar{\varepsilon}_1$. The update law of the RBFNN weights is designed as

$$\dot{\hat{w}} = -\Gamma Sh \tag{26}$$

Bringing 26 into 25 yields

$$\begin{aligned}\dot{V}_3 &= -k_2S^2 - k_3|S|^{\gamma_1+1} - S\varepsilon_1\\ &\le -k_2S^2 - k_3|S|^{\gamma_1+1} + |S|\,\bar{\varepsilon}_1\\ &= -k_2S^2 - |S|\left(k_3|S|^{\gamma^1} - \bar{\varepsilon}_1\right)\end{aligned} \tag{27}$$

when $k_3 > \bar{\varepsilon}/|S|^{\gamma_1}$, $k_3|S|^{\gamma_1} - \bar{\varepsilon} = \varepsilon_2$, where $\varepsilon_2 > 0$, thus:

$$\begin{aligned}\dot{V}_3 &\le -2k_2V_2 - \varepsilon_2\,|S| \le -2k_2V_2 - \sqrt{2}\varepsilon_2 V_2^{1/2}\\ &< -\alpha_1 V_2^{1/2} - \beta_1 V_2 < 0\end{aligned} \tag{28}$$

According to **lemma 1**, V_2 can converge to 0 in finite time.

The control input u_c in formula 24 is the unconstrained, to prevent the control input saturation, define $u_d = u_c$, where u_d is the desired value in the next step, and the state error $e_3 = u_d - u_{con}$. u_{con} satisfies the constrained control input of the saturation function *tanh*; therefore, parameter u_f must exist, such that $u_{con} = u_{\max}\tanh(u_f/u_{\max})$, where u_{max} is the maximum input.

$$\dot{u}_{con} = \left(1 - \tanh^2(u_f/u_{\max})\right)\dot{u}_f \tag{29}$$

Step 4: Establish the Lyapunov function $V_4 = \frac{1}{2}e_3^2$ and derive V_3 and bring it into 29 to obtain:

$$\begin{aligned}\dot{V}_4 &= e_3\dot{e}_3\\ &= e_3(\dot{u}_{\mathrm{d}} - \dot{u}_{con})\\ &= e_3\left(\dot{u}_d - \left(1 - \tanh^2(u_f/u_{\max})\right)\dot{u}_f\right)\end{aligned} \tag{30}$$

$\dot{u}_f$ is designed as

$$\dot{u}_f = \begin{cases}\left(k_4e_3 + |e_3|^{\gamma_2}sgn(e_3) + \dot{u}_d\right) / \left(1 - \tanh^2(u_f/u_{\max})\right), \delta \ge \Delta\\ |\delta e_3|^{\gamma_2}sgn(e_3) + \dot{u}_d / \left(1 - \tanh^2(u_f/u_{\max})\right), \delta < \Delta\end{cases} \tag{31}$$

where $\delta{=}\left|u_f\right|{-}2u_{\max}$, Δ is a smaller normal value. $\gamma_2 \in (0, 1)$. The convergence of the controller is discussed in the following cases. When $\delta \ge \Delta$, substituting 31 into 30 yields

$$\begin{aligned}\dot{V}_4 &= -k_4e_3^2 - |e_3|^{\gamma_2+1} = -2k_4V_3 - 2^{(\gamma_2+1)/2}V_3^{(\gamma_2+1)/2}\\ &< -\alpha_2 V_4^{(\gamma_2+1)/2} - \beta_2 V_4\end{aligned} \tag{32}$$

where $0 < \alpha_2 < 2^{(\gamma_2+1)/2}$, $2k_3 = \beta_2$. According to **Lemma 1**, V_4 can converge to 0 in finite time. When $\delta < \Delta$, substituting 31 into 30 yields

$$\begin{aligned} \dot{V}_4 &= -\left(|\delta|^{\gamma_2}|e_3|^{\gamma_2+1}\right)/\left(1-\tanh^2(u_c/u_{\max})\right) \\ &= -\left(|\delta|^{\gamma_2}2^{(\gamma_2+1)/2}/\left(1-\tanh^2(u_c/u_{\max})\right)\right)V_4^{\alpha_3} \\ &= -cV_4^{\alpha_3} \end{aligned} \tag{33}$$

where $\alpha_3 = (\gamma_2+1)/2, c = |\delta|^{\gamma_2}2^{(\gamma_2+1)/2}/\left(1-\tanh^2(u_c/u_{\max})\right)$, and $\tanh(u_c/u_{\max}) < 1$, so $c > 0$. According to **Lemma 2**, V_4 can converge in finite time.

LEMMA 2: Chu et al. (2022) Suppose that there is a positive definite continuous Lyapunov function $V(x,t)$ defined on $U_1 \times R^+$, where $U_1 \subseteq U \subseteq R_{\text{n}}$. R_n is a neighborhood of the origin, and $V(x,t) \le -cV^{\alpha}(x,t), \forall x \in U_1 \backslash \{0\}$, where $c > 0, 0 < \alpha < 1$. Then, the origin of the system is locally finite time stable. The settling time $T \le V^{1-\alpha}\left(x(t_0),t_0\right)/c(1-\alpha)$ satisfies for a given initial condition $x(t_0) \in U_1$.

3.2. Human decision search algorithm

The human decision search algorithm (HDSA) is a swarm optimization technique that mimics the decision-making process of a human crowd. In many post-apocalyptic survival games or films, the strong group consciousness of humans is often portrayed, but the importance of individual consciousness is also emphasized. In human groups, a small group of individuals called decision-makers make the final decisions based on their experience and personal status. However, the decision of the decision-maker is not necessarily optimal. When the number of individuals in the group is small, it is important to involve more people in the decision-making process to guide the development of the group and to avoid the excessive impact of individual decisions on the group. However, when the number of individuals in the group is large, the proportion of decision-makers should be reduced and only a few elite individuals should be selected to determine the development of the group. This is because too many people involved in the decision-making process may take more time, and the experience of ordinary people may not be as good as that of elite individuals. Because people have emotions, they can think both rationally and emotionally when dealing with problems, and these two opposing ways of thinking must coexist.

Apart from the decision-makers, the rest of the human population is referred to as the executors, consisting of individuals who have no or less ability to make decisions. They carry out the optimal decisions made by the decision-makers. However, individuals among the executors who have some decision-making ability should be encouraged to seek more humane decisions based on the optimal decisions. These decisions should become more adapted to the current environment over time. The number of decision-makers is fixed, and elite individuals in the human population will always be selected as decision-makers. Over time, any individual has the potential to become a decision-maker, and the current decision-maker may become an executor.

In a human population, there are always individuals who question the current decision or believe they have a better one, including the decision-makers themselves. These individuals are known as adventurers, and their numbers and identities are random, making them a source of uncertainty within the population. Although adventurers can lead people to a better life, they can also lead them to disaster. Adventurers, on the other hand, inherit the current optimal choices of the human population and take them into account when making decisions. However, more adventurous individuals will also seek out possible optimal decisions based on their own state. To avoid harming the human population, adventurers must consider whether the decisions they make are more beneficial to their own survival. Additionally, there is a chance that an adventurer will become a decision-maker if they come up with a better or suboptimal decision. Based on the above analysis, the proposed algorithm for optimizing the human decision population consists of three main components: decision updating for decision-makers, decision updating for executors, and decision updating for adventurers.

3.2.1. Decision updates for decision makers

The number of decision-makers is fixed in proportion to the total number of people, and the number of decision-makers is 20–50% of the total number of people. The decision-makers make their decisions based on individual experience as well as individual characteristics. The sine and cosine functions are used to distinguish between rational and emotional decisions by people, and the individuals are randomly updated due to the random adoption of rational and emotional decisions by people.

$$x_i^{t+1} = \begin{cases} r_1 x_i^t \sin\left(r_2\left|r_3 x_{ibest}^t - x_i^t\right|\right), R < 0.5 \\ r_1 x_i^t \cos\left(r_2\left|r_3 x_{ibest}^t - x_i^t\right|\right), R \ge 0.5 \end{cases} \tag{34}$$

where x_i^t denotes the t_{th} iteration of the i_{th} human individual. r_1 is a non-linear term, $r_1 = 2*\left(1 - i/(\alpha_1 * d_{num})\right)$. d_{num} is the number of decision-makers. α_1 is a random number between $(0,1)$. $r_2 = \alpha_2 2\pi$ and α_2 is the random number between $(0,1)$. $r_3 = 2\alpha_3$, α_3 is a random number between $(0,1)$. r is the random number between $(0,1)$. x_{ibest}^t is the individual optimal solution for 1 to t iterations.

3.2.2. Decision updates for executors

Except for the decision-maker, the rest of the individuals are the executors. Among the executors, individuals with a fitness that is higher than the intermediate fitness are ordinary executors that must follow the optimal decision of the decision-maker. Individuals with a fitness below the intermediate fitness are considered as executors with some decision-making ability, and this group can continue to explore the next optimal decision that may exist based on the current optimal decision.

$$x_i^{t+1} = \begin{cases} x_{best}^t + \beta_1\left|\left(x_i^t - x_m^t\right)/\left(f_i^t - f_m^t\right)\right|, f_i^t > f_m^t \\ sgn(x_e^t)exp\left(\left|x_{best}^t - x_i^t\right|/\beta_2\right), f_i^t \le f_m^t \end{cases} \tag{35}$$

where x_{best}^t is the current global best individual and x_{worst}^t is the current global worst individual. $x_e^t = x_{best}^t - x_{worst}^t$. f_i^t is the fitness of the i_{th} individual, $f_m^t = \left(f_{best}^t + f_{worst}^t\right)/2$, f_{best}^t is the current best fitness, and f_{worst}^t is the current worst fitness. β_1 is the random number of normal distribution with mean 0 and variance 1. The sgn function determines the direction of exploration of individuals. $\beta_2{=}t^2/f^t_{best}$ indicates that a more favorable decision result can be obtained over time.

3.2.3. Decision updates for adventurers

The adventurers are random individuals and the number of adventurers is also random. If the adventurer's fitness is less than the average fitness, the adventurer randomly explores based on the current optimal solution. If the adventurer's fitness is higher than the average fitness, the adventurer will continue to explore in the optimal direction according to the current state of the individual.

$$x_i^{t+1} = \begin{cases} x^t_{best} + c_1 \left| x^t_{best} - x^t_i \right|, f^t_i > f^t_{avr} \\ x^t_i + (2c_2 - 1) \left|\left| x^t_e \right|\right|_2 sgn(x^t_e), f^t_i \le f^t_{avr} \end{cases} \tag{36}$$

where c_1 is a normally distributed random number with mean 0. c_2 is a random number between $(0, 1)$ with variance 1. $\|x^t_e\|_2$ is the Euclidean norm of x^t_e and f^t_{avr} is the current mean fitness.

Based on the above discussion, the proposed HDSA has three steps. The first step performs a global random search using the formula 34. In the second step, a local search is performed based on the first step using the formula 35. The third step performs a second global random search using the formula 36 on the basis of the first and second steps. HDSA framework as Algorithm 1.

3.3. Yaw controller and linear velocity controller

According to the control algorithm in the "RBFNN-Based Active Fault-Tolerant Control Algorithm" section, the AFTC is used to design controllers in this section to follow the desired yaw angle ψ_d and desired linear velocity v_d. The robot linear velocity sliding mode surface is: $S_v = \alpha_v e_v + \beta_v {e_v}^{\lambda_v}$, where $e_v = v_d - v$. The sliding mode convergence law is $\dot{S}_v = -k_{2v} S_v - k_{3v} |S|^{\gamma_{1v}} \text{sgn}(S_v)$.

The proof of convergence for the velocity controller is similar to that for the general-purpose controller in the "RBFNN-Based Active Fault-Tolerant Control Algorithm" section. The unconstrained control law is designed as

$$F_{uc} = m \left(\dot{v}_d - \hat{w}^T_v h_v + k_{2v} S_v + k_{3v} |S_v|^{\gamma_{1v}} \text{sgn}(S_v) \right) \tag{37}$$

The anti-input saturation controller of linear velocity is designed as

$$\begin{cases} F_{uf} = \begin{cases} \int \left(k_{4v} e_F + |e_F|^{\gamma_{2v}} sgn(e_F) + \dot{F}_{uc} \right) \\ \quad / \left(1 - \tanh^2(F_{uf}/F_{\max}) \right) dt \quad , \delta_v \ge \Delta_v \\ \int |\delta_v e_F|^{\gamma_{2v}} sgn(e_F) + \dot{F}_{uc} / \left(1 - \tanh^2(F_{uf}/F_{\max}) \right) dt \\ \qquad\qquad , \delta_v < \Delta_v \end{cases} \\ F_{ucon} = F_{\max} \tanh(F_{uf}/F_{\max}) \end{cases} \tag{38}$$

Where $e_F = F_{uc} - F_{ucon}$.

The yaw angle controller is $\omega_d = k_\psi e_\psi + \dot{\psi}_d$, where $e_\psi = \psi_d - \psi$. The yaw angle sliding mode surface is $S_\omega = e_\omega + \alpha_\psi e_\psi + \beta_\psi {e_\psi}^{\lambda_\psi}$. The sliding mode convergence law is $\dot{S}_\omega = -k_{2\omega} S_\omega - k_{3\omega} |S_\omega|^{\gamma_{1\omega}} \text{sgn}(S_\omega)$.

```
Input: input parameters $M$, $d_n$, $pop$, $dim$, $l_b$, and $u_b$
Output: output $x_{best}$, $f_{min}$
Initialize individuals, constrain the upper and
  lower bounds of individuals, calculate the
  individual fitness, initialize the global
  optimal solution $x_{best}$ and the optimal fitness
  $f_{min}$;
while $t < M$ do
    Find the current global optimal solution $x_{best}$
      and the current individual optimal solution
      $x_{ibest}$  $r = rand(1)$
    for $i : d_n$ do
        Use 34 to update the decision-maker's
          decision and calculate the individual
          fitness
    Find the current global optimal solution $x_{best}$
      and the global worst solution $x_{worst}$, compute
      the intermediate solution and intermediate
      solution fitness.
    for $d_n + 1 : pop$ do
        Use 35 to update the executor's decision and
          calculate the individual fitness
    Calculate the average fitness, randomly select
      $a_n$ individuals
    for $1 : a_n$ do
        Use 36 to update the adventurer's decision
          and calculate the individual fitness
    Find the current global optimal solution $x_{best}$,
      the individual optimal solution $x_{ibest}$, and the
      optimal fitness $f_{min}$.
    $t = t + 1$
return $x_{best}$, $f_{min}$
```

Algorithm 1. HDSA.

The unconstrained control law is designed as

$$T_{rc} = I \left(\dot{\omega}_d - \hat{w}^T_\omega h_\omega + k_{2\omega} S_\omega + k_{3\omega} |S_\omega|^{\gamma_{1\omega}} \text{sgn}(S_\omega) \right) \tag{39}$$

The anti-input saturation controller of the yaw angle is designed as

$$\begin{cases} T_{rf} = \begin{cases} \int \left(k_{4\omega} e_T + |e_T|^{\gamma_{2\omega}} sgn(e_T) + \dot{T}_{rc} \right) \\ \quad / \left(1 - \tanh^2(T_{rf}/T_{\max}) \right) dt \quad , \delta_\omega \ge \Delta_\omega \\ \int |\delta_\omega e_T|^{\gamma_{2\omega}} sgn(e_T) + \dot{T}_{rc} / \left(1 - \tanh^2(T_{rf}/T_{\max}) \right) dt, \\ \quad \delta_\omega < \Delta_\omega \end{cases} \\ T_{rcon} = T_{\max} \tanh(T_{rf}/T_{\max}) \end{cases} \tag{40}$$

where $e_T = T_{rc} - T_{rcon}$. The controller parameters are not described in this section as they have been discussed in the "RBFNN-Based Active Fault-Tolerant Control Algorithm" section.

The input to the angular velocity neural network is both the yaw error and the angular velocity error, and the output is the

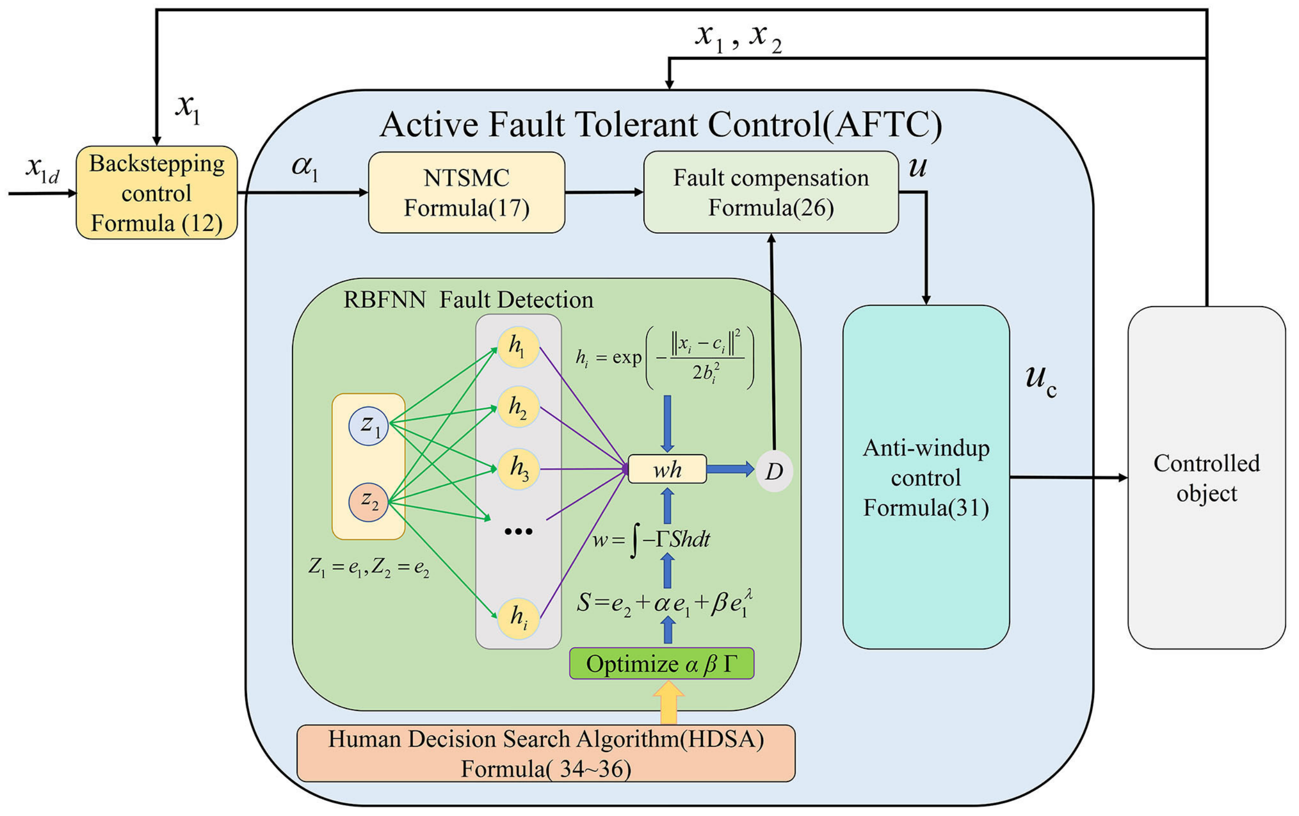

x_1, x_2
x_1
x_{1d}
Backstepping control Formula (12)
α_1
Active Fault Tolerant Control(AFTC)
NTSMC Formula(17)
Fault compensation Formula(26)
u
RBFNN Fault Detection
$h_i = \exp\left(-\frac{\|x_i - c_i\|^2}{2b_i^2}\right)$
z_1
z_2
h_1
h_2
h_3
h_i
wh
D
$Z_1 = e_1, Z_2 = e_2$
$w = \int -\Gamma S h dt$
$S = e_2 + \alpha e_1 + \beta e_1^{\lambda}$
Optimize $\alpha\ \beta\ \Gamma$
Human Decision Search Algorithm(HDSA) Formula(34~36)
u_c
Anti-windup control Formula(31)
Controlled object

FIGURE 6
AFTC framework.

uncertainty term in the angular velocity control. The coordinate vector matrix of the centroids of the Gaussian basis function neurons in the angular velocity neural network is $c_\psi = \begin{bmatrix} -1.6 & -0.8 & -0.4 & -0.2 & -0.1 & 0 & 0.1 & 0.2 & 0.4 & 0.8 & 1.6 \\ -1.6 & -0.8 & -0.4 & -0.2 & -0.1 & 0 & 0.1 & 0.2 & 0.4 & 0.8 & 1.6 \end{bmatrix}_{2*11}$

The width of the Gaussian basis function $b_\psi = 0.1, i = 1 \cdots 11$.

The input to the linear velocity neural network is the velocity error and the output is the linear velocity control uncertainty term. The coordinate vector matrix of the centroids of the Gaussian basis function of the neurons in the linear velocity neural network is $c_v = \begin{bmatrix} -1.6 & -0.8 & -0.4 & -0.2 & -0.1 & 0 & 0.1 & 0.2 & 0.4 & 0.8 & 1.6 \end{bmatrix}_{1*11}$. The width of the Gaussian basis function $b_v = 0.1, i = 1 \cdots 11$.

Based on the above discussion, the proposed framework for the AFTC is shown in Figure 6.

4. Simulation results

In the section entitled "HDSA's Related Work", we have demonstrated the advantages of the proposed HDSA; therefore, in this section, the HDSA is used to optimize the sliding mode surface parameters of the yaw controller and the linear velocity controller. As the weight update parameters of the RBFNNs are related to the sliding mode parameters, this also indirectly optimizes the RBFNNs.

The parameters to be optimized for yaw angle control are the sliding mode surface coefficients α_ω , β_ω and the neural network update coefficient Γ_ω. According to the idea of AFTC, the presence of $-3N.m$ of disturbance torque in the robot model simulates the worst case. The initialized optimization algorithm parameters are as follows: dimension is 3, the number of populations is 20, the number of max iterations is 10, and the upper limit of parameters is 20 and the lower limit is -20.

The evaluation function of the yaw controller is designed as $f_{obj} = 0.8 * |e_\psi| + 0.1 * |e_\omega| + 0.01 * |T_{rc}|$. For yaw control, we want to reduce both the yaw error and the yaw velocity error with the smallest control input. As the control objective is to eliminate the yaw error, the yaw error is given the largest weight in the evaluation function. To keep the control input and yaw error in the same order, the control input weight is reduced. The optimization parameters for the yaw controller are shown in Figure 7.

As shown in Figure 7, the optimized parameters converge after eight iterations. The values of $\Gamma_\omega = 20$, $\alpha_\omega = 7.4407$, and $\beta_\omega = 2.9369$ are obtained through the optimization process.

The optimized parameters are substituted into the AFTC and the control results are compared with the unoptimized AFTC, NTSMC, and SMC. Before 10 s, the yaw angle is influenced by a torque with a mean value of $-1N.m$ and a mean square error of 0.1. After 10 s, the yaw angle is influenced by a torque with a mean value of $-3N.m$ and a mean square error of 0.1. The control parameters are given in Table 3.

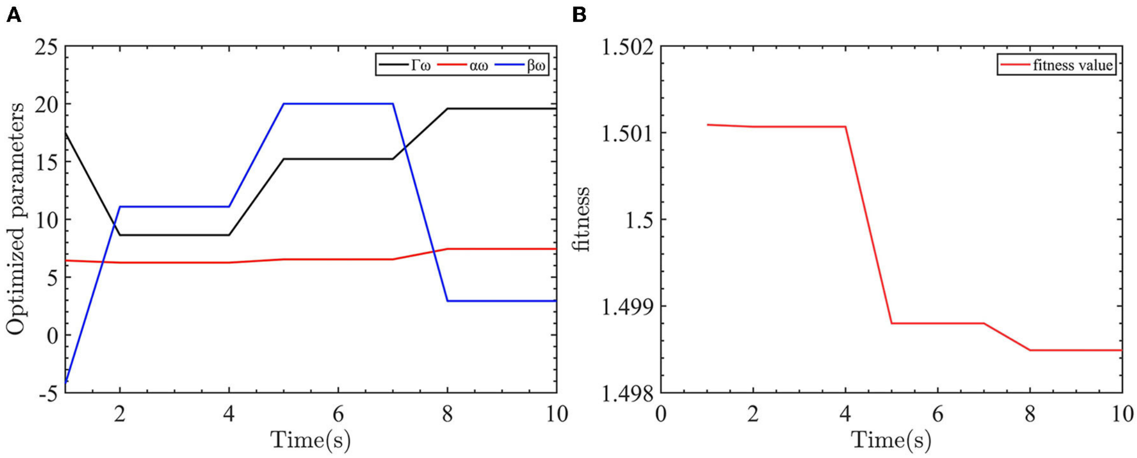


FIGURE 7
Yaw control parameter optimization and fitness of the yaw controller objective function. **(A)** The optimized parameters of yaw controller. **(B)** The objective function output value.

TABLE 3 Parameters of yaw angle controllers.

Controllers	Parameters	Value
Proposed AFTC	$k_{1\psi}$	2
	$\alpha_\omega, \beta_\omega, \lambda_\omega$	1, 2, 3
	$k_{2\omega}, k_{3\omega}, \gamma_\omega$	1, 5, 0.5
	$k_{4\omega}$	5
	Γ_ω	10
NTSMC	$k_{1\psi}$	2
	$\alpha_\omega, \beta_\omega, \lambda_\omega$	1, 2, 3
	$k_{2\omega}, k_{3\omega}, \gamma_\omega$	5, 20, 0.5
SMC	$k_{1\psi}$	2
	$k_{2\omega}, k_{3\omega}$	5, 5

The results of the yaw angle controller are shown in Figure 8.

In Figure 8A, the optimized AFTC has a significantly faster response speed (pink line). Despite being influenced by a $-1\ N.m$ torque disturbance in the range of 0–10 s, the AFTC, NTSMC (green line), and SMC (red line) maintain their robustness and are not affected by the disturbance. After 10 s, the yaw angle is subjected to a torque of $-3N.m$, in which case reliance on the robustness of the controller can no longer guarantee yaw angle control performance, as shown in the 10–11 s enlargement in Figure 8A. The SMC is unable to follow the desired yaw angle with a static error of ~0.05 *rad*, and the NTSMC also has a small static difference.

As shown in Figure 8B, the proposed AFTC (pink line) and the optimized AFTC (orange line) do not enter the driver saturation state. The NTSMC (purple line) and the SMC (green line) enter the driver saturation state. Compared with the conventional SMC (green line) and NTSMC (purple line) control inputs, which have high-frequency input chatter, the control input of the proposed AFTC is more stable. This suggests that the robustness achieved by the conventional SMC comes at the expense of control input performance. In Figure 8C, the output of the radial basis function neural network (RBFNN) is displayed, showing a value of 1 before 10 s and 3 after 10 s. The RBFNN can estimate the unknown yaw disturbances online. The RBFNN weights are updated accordingly, as shown in Figure 8.

The parameters to be optimized for the velocity controller are the sliding mode surface coefficients α_v and β_v and the neural network update coefficients Γ_v. The presence of $-5N$ force in the robot model simulates the worst case. The initialized optimization algorithm parameters are as follows: the dimension is 3, the number of populations is 20, the number of maximum iterations is 10, and the upper limit of parameters 20 and the lower limit is 20.

The evaluation function is designed as $f_{obj} = 0.8 * |e_v| + 0.02 * |F_{uc}|$. When controlling the linear velocity, we want to minimize the linear velocity error with the smallest control input. Therefore, the linear velocity error has the largest weight in the evaluation function. The weight of the control input is reduced to keep the control input and the linear velocity error at the same level. The linear velocity controller optimization parameters are shown in Figure 9.

As shown in Figure 9, the optimization parameters converge after two iterations. The optimized parameters are $\Gamma_v = 15.6467$, $\alpha_v = 16.1866$, and $\beta_v = 20$.

These parameters are used in the proposed AFTC, and the control results are compared and analyzed with the unoptimized AFTC, NTSMC, and SMC controllers. Before 10 s, the linear velocity is affected by a force with a mean value of $-2N$ and a mean square error of 0.1. After 10 s, the velocity is influenced by a force with a mean value of $-5N$ and a mean square error of 0.1. The velocity controller parameters are given in Table 4.

The control results of linear velocity controllers are shown in Figure 10.

Similar to the performance of the yaw control, in Figure 10A, the optimized AFTC (pink line) responds faster compared with the proposed AFTC (purple line) and SMC (red line). Between 0 and 10 s, when the line speed is subjected to -2N force, AFTC (purple line), NTSMC (green line), and SMC (red line) are not affected

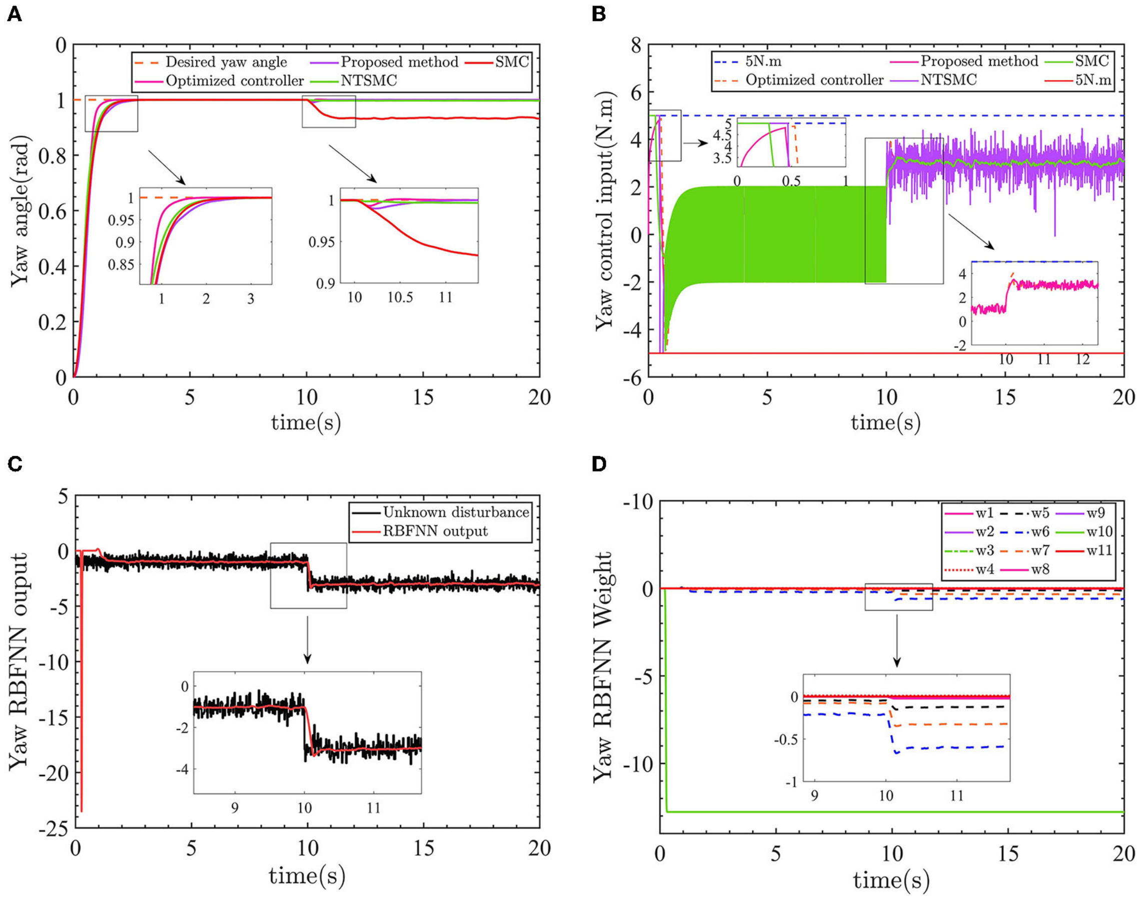


FIGURE 8

(A) The yaw angle control results. **(B)** Control input torque. **(C)** Yaw angle RBFNN output value. **(D)** Yaw angle RBFNN weight.

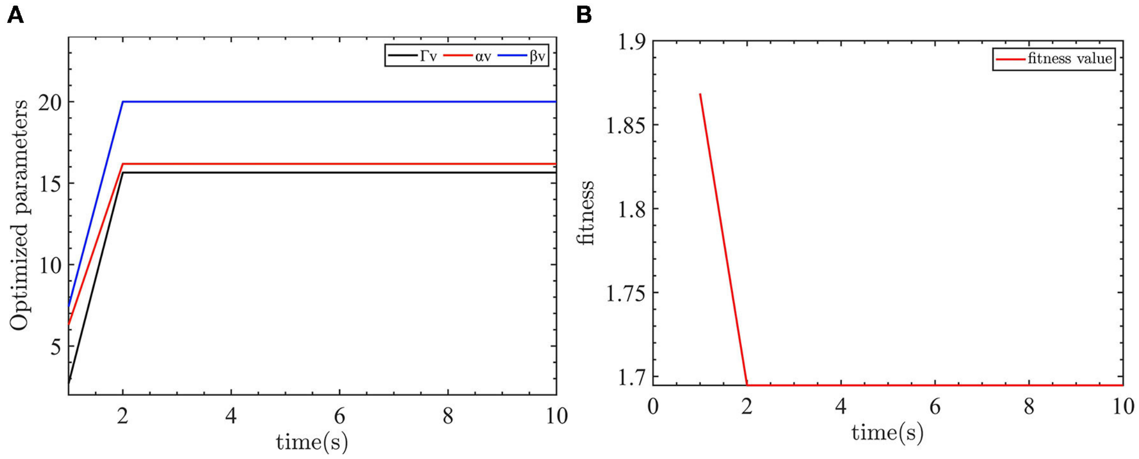


FIGURE 9

Velocity control parameter optimization and fitness of the velocity controller objective function. **(A)** The optimized parameters of velocity controller. **(B)** The objective function output value.

by the disturbances. After 10 s, the linear velocity is subjected to a force of $-5N$ and the velocity control performance cannot be guaranteed by the NTSMC and SMC. There is a static error of $\sim 0.05m/s$ for the NTSMC and $\sim 0.6m/s$ for the SMC, as shown in the 9–12 s enlargement in Figure 10A. Both the proposed AFTC and the optimized AFTC can follow the desired linear velocity, and the velocity controller is almost unaffected by the $-5N$ force using the optimized parameters. The proposed AFTC and the optimized AFTC can effectively track the desired linear velocity, with minimal impact from the $-5N$ force disturbance. The velocity controller of the AFTC is almost unaffected by the disturbance, indicating its robustness and ability to maintain precise control performance.

The previous discussion has highlighted the improved responsiveness and robustness of the optimized AFTC. To further emphasize the advantages of the optimized AFTC, the output value of the evaluation function is used as a criterion to evaluate the performance of the four controllers. A smaller output value of the evaluation function indicates better controller performance. The output values of the evaluation functions for the four controllers are depicted in Figure 11.

As shown by the green lines in Figures 12A, B, the optimized AFTC controller exhibits the smallest value of the evaluation function. This signifies that the optimized AFTC achieves the best performance among the four controllers. As the linear velocity and yaw angle are consistently subjected to external disturbances, the output value of the evaluation function continually increases. This is because of the fact that the control inputs are not equal to zero. In the case of large external disturbances, the NTSMC and SMC controllers can no longer eliminate the yaw angle error and the linear velocity error. Consequently, the output value of the evaluation function rapidly increases, as indicated by the red and blue lines.

To further verify the effectiveness of the proposed algorithm, the AFTC is used to design the yaw angle controller and the velocity controller. The desired yaw angle and the desired linear velocity is planned by the LOS algorithm. The optimized parameters are selected as the controller's parameters. The LOS algorithm

TABLE 4 The parameters of velocity controllers.

Controllers	Parameters	Value
Proposed AFTC	$\alpha_v, \beta_v, \lambda_v$	1, 2, 3
	k_{2v}, k_{3v}, γ_v	1, 5, 0.5
	k_{4v}	5
	Γ_v	10
NTSMC	$\alpha_v, \beta_v, \lambda_v$	1, 2, 3
	k_{2v}, k_{3v}, γ_v	5, 20, 0.5
SMC	k_{2v}, k_{3v}	5, 5

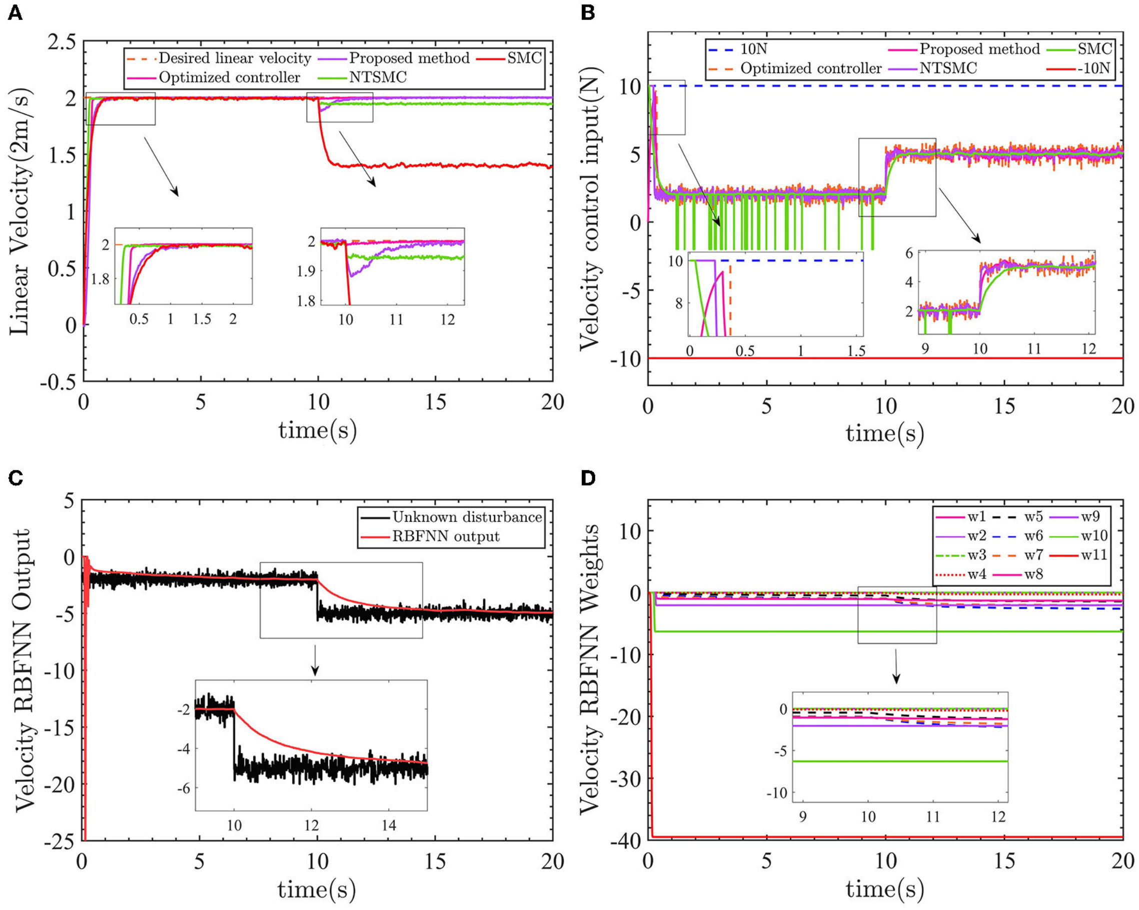


FIGURE 10
Linear velocity control results. **(A)** Velocity control results. **(B)** Control input force. **(C)** Velocity RBFNN output value. **(D)** Velocity RBFNN weight.

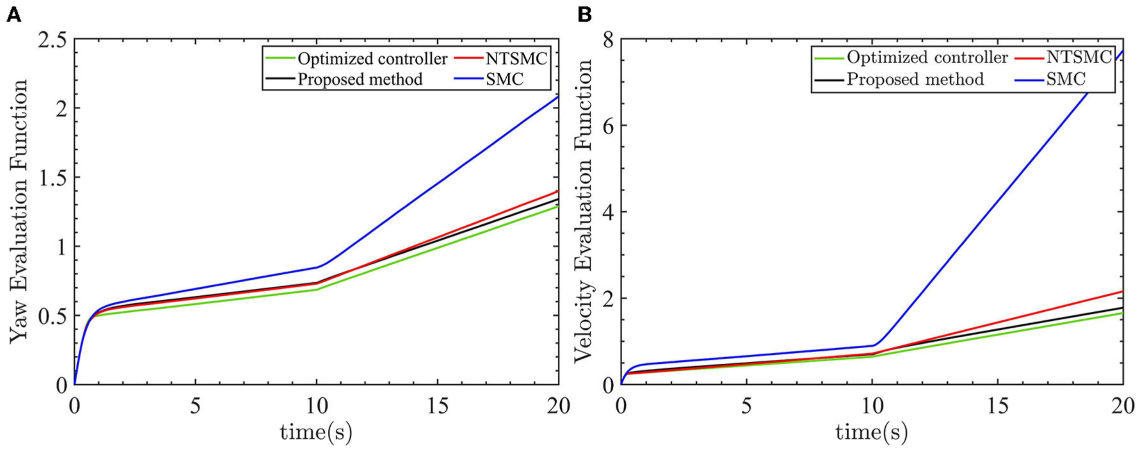


FIGURE 11
Four control evaluation function outputs. **(A)** Yaw angle evaluation function outputs. **(B)** Velocity evaluation function outputs.

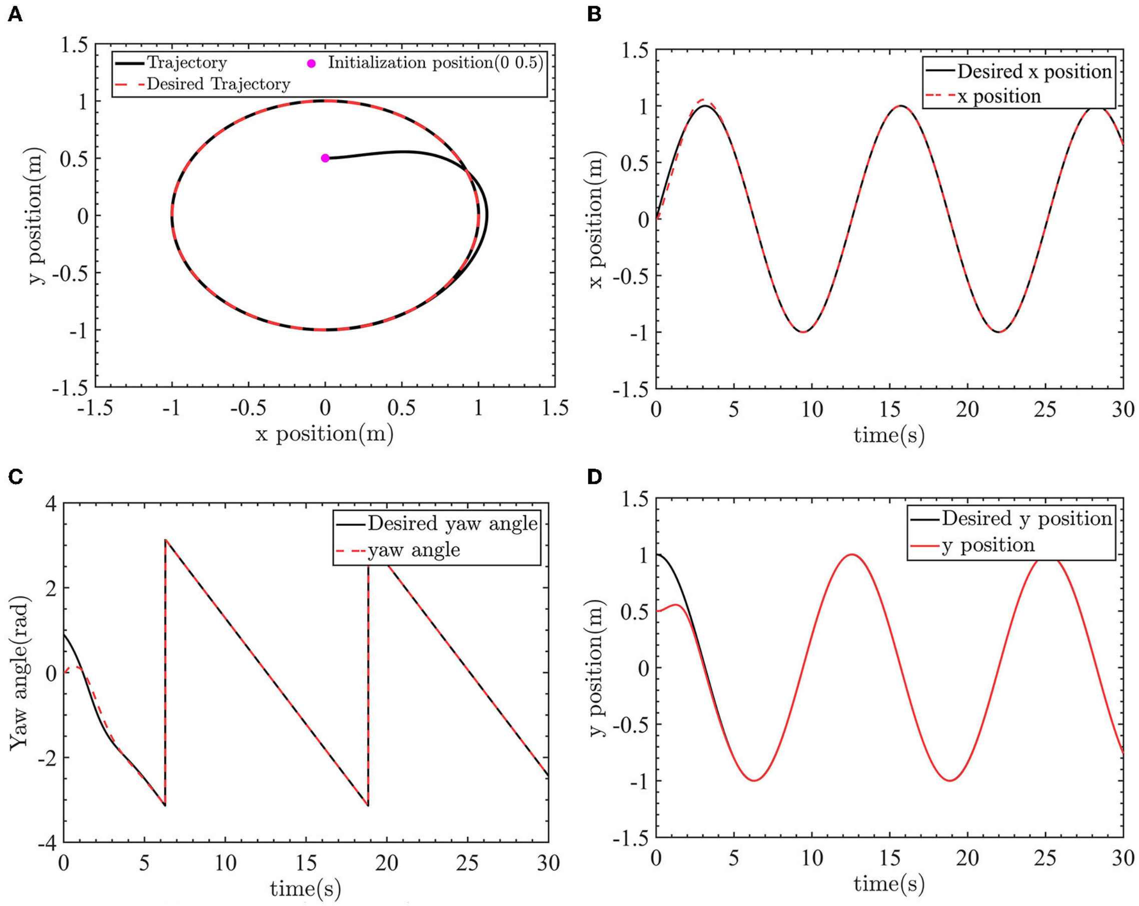


FIGURE 12
The robot tracks the desired trajectory. **(A)** Tracking the circle desired trajectory. **(B)** X-position control. **(C)** Yaw angle control. **(D)** Y-position control.

and the improved LOS algorithm can be found in the author's previous work (Wang et al., 2022b). The desired trajectory is a circular trajectory with radius $R = 1m$, angular velocity $\omega_r = 0.5rad/s$, and linear velocity $v_r = 0.5m/s$. The initial position and pose of the robot is $\left[0\text{m}, 0.5\text{m}, 0\text{rad}\right]$. A drag force of $-2N$ and a torque of $-1N.m$ are applied to the robot. The

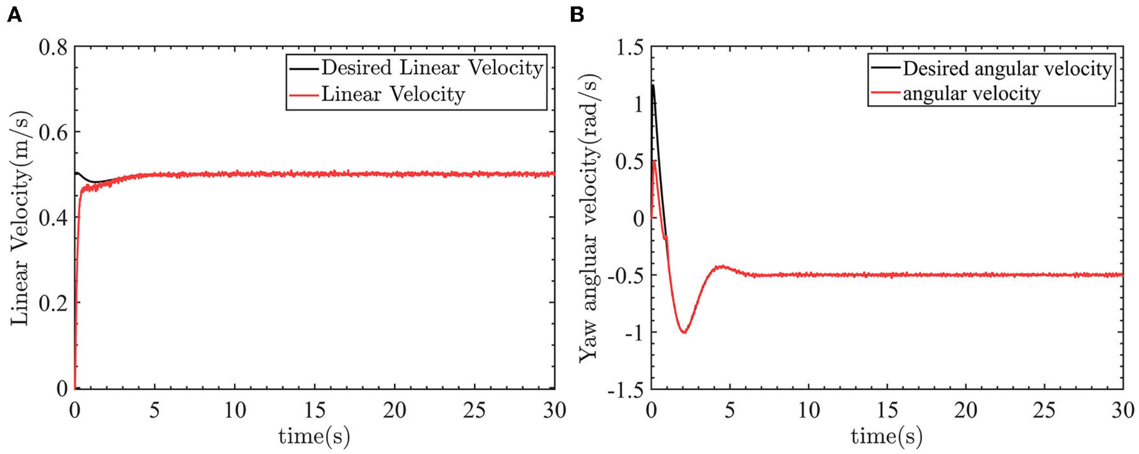


FIGURE 13
The control results of linear velocity and yaw angular velocity. **(A)** Linear velocity control. **(B)** Yaw angle velocity control.

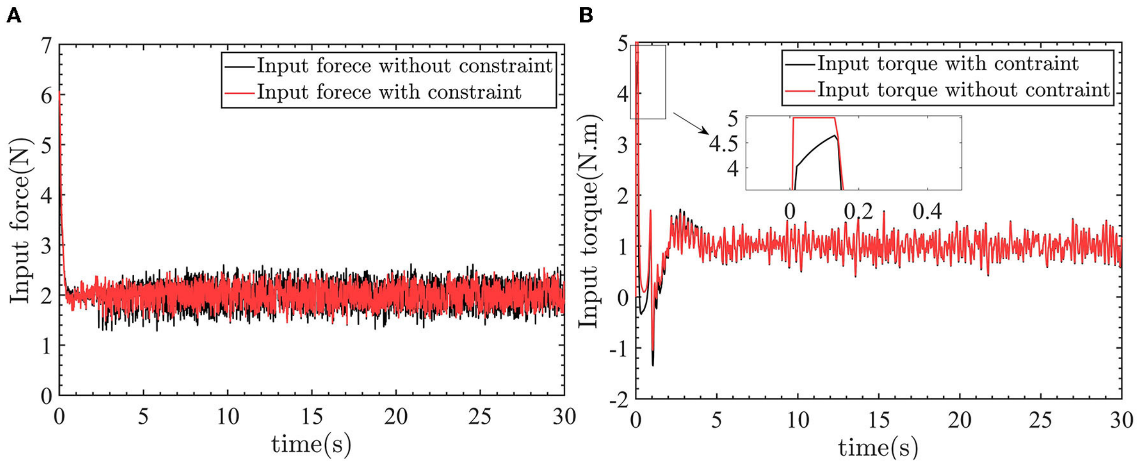


FIGURE 14
The linear velocity control input and yaw angular velocity control input. **(A)** Control input force. **(B)** Control input torque.

LOS algorithm is

$$\begin{cases} \psi_L{=}\psi_r - \alpha \\ \alpha = \arctan(e_y/\Delta) \\ v_L{=}v_r + ke_x \end{cases} \tag{41}$$

where ψ_L, v_L are the desired yaw angle and desired linear velocity planned by the LOS algorithm. e_x, e_y is the position error in Frenet-Serret (F-S) frame. Δ and k are the positive adjustable parameters.

The control results of the robot tracking the desired circle trajectory are shown as Figures 12–14. The robot position control and yaw angle control are shown in Figure 12.

The robot can track the desired trajectory. The actual position pose of the robot is consistent with the desired position pose. The linear velocity control and angular velocity control are shown in Figure 13.

In Figure 13A, the linear velocity can track the desired linear velocity of $0.5m/s$. In Figure 13B, the angular velocity can track the desired angular velocity of $-0.5rad/s$. Figure 14 shows the linear velocity control input and yaw angle velocity control input.

TABLE 5 The single-peak test functions.

Function	Initial range	Fmin
$f_1(x) = \sum_{i=1}^{30} x_i^2$	$-100 \leq x_i \leq 100$	0
$f_2(x) = \sum_{i=1}^{30} \|x_i\| + \prod_{i=1}^{30} \|x_i\|$	$-10 \leq x_i \leq 10$	0
$f_3(x) = \sum_{i=1}^{30} \left(\sum_{j=1}^{i} x_j \right)^2$	$-100 \leq x_i \leq 100$	0
$f_4(x) = \max_i \{\|x_i\| \; 1 \leq i \leq 30\}$	$-100 \leq x_i \leq 100$	0
$f_5(x) = \sum_{i=1}^{29} \left[100\left(x_{i+1} - x_i^2\right)^2 + (x_i - 1)^2 \right]$	$-30 \leq x_i \leq 30$	0
$f_6(x) = \sum_{i=1}^{29} (\|x_i + 0.5\|)^2$	$-100 \leq x_i \leq 100$	0
$f_7(x) = \sum_{i=1}^{30} ix_i^4 + random\,[0, 1)$	$-1.28 \leq x_i \leq 1.28$	0

TABLE 6 The multi-peak test functions.

Function	Initial range	Fmin
$f_8(x) = -\sum_{i=1}^{30} \left(x_i \sin\left(\sqrt{\lvert x_i \rvert}\right)\right)$	$-500 \le x_i \le 500$	-12569.5
$f_9(x) = \sum_{i=1}^{30} \left[x_i^2 - 10\cos\left(2\pi x_i + 10\right)\right]$	$-5.12 \le x_i \le 5.12$	0
$f_{10}(x) = -20exp^{\left(-0.2\sqrt{\frac{1}{30}\sum_{1}^{30} x_i^2}\right)} - exp^{\left(\frac{1}{30}\sum_{1}^{30}\cos 2\pi x_i\right)} + 20 + c$	$-100 \le x_i \le 100$	0
$f_{11}(x) = \frac{1}{4000}\sum_{i=1}^{30} x_i^2 - \prod_{i=1}^{30}\cos\left(\frac{x_i}{\sqrt{i}}\right) + 1$	$-600 \le x_i \le 600$	0
$f_{12}(x) = \frac{\pi}{30}\left\{100\sin^2\left(\pi y_1\right) + \sum_{i=1}^{29}\left(y_i - 1\right)^2 \times\right.$ $\left.\left[1 + 10\sin^2\left(\pi y_{i+1}\right)\right] + \left(y_n - 1\right)^2\right\} + \sum_{i=1}^{30} u\left(x_i, 10, 100, 4\right)$	$-50 \le x_i \le 50$	0
$f_{13}(x) = 0.1\left\{\sin^2\left(\pi 3x_1\right) + \sum_{i=1}^{29}\left(x_i - 1\right)^2\left[\sin^2\left(3\pi x_{i+1}\right)\right] +\right.$ $\left(x_n - 1\right)^2\left[1 + \sin^2\left(2\pi x_{30}\right)\right] + \sum_{i=1}^{30} u\left(x_i, 5, 100, 4\right)$	$-50 \le x_i \le 50$	0

TABLE 7 The fixed-dimensional multi-peak test functions.

Function	Initial range	Fmin
$f_{14}(x) = \left[\frac{1}{500} + \sum_{j=1}^{25}\frac{1}{j + \sum_{i=1}^{2}\left(x_i - a_{ij}\right)^6}\right]^{-1}$	$-65.536 \le$ $x_i \le 65.536$	1
$f_{15}(x) = \sum_{i=1}^{11}\left[a_i^2 - \frac{x_1\left(b_i^2 + b_i x_2\right)}{b_i^2 + b_i x_3 + x_4}\right]$	$-5 \le x_i \le 5$	0.0003075
$f_{16}(x) = 4x_1^2 - 2.1x_1^4 - \frac{1}{3}x_1^6 + x_1x_2 - 4x_2^2 + 4x_2^4$ $f_{17}(x) = \left(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2$ $+10\left(1 - \frac{1}{8\pi}\right)\cos x_1 + 10$	$-5 \le x_i \le 5$ $-5 \le x_1 \le 10$ $0 \le x_2 \le 15$	-1.0316 0.398
$f_{18}(x) = \left[1 + \left(x_1 + x_2 + 1\right)^2\right. \times$ $\left.\left(19 - 14x_1 + 3x_1^2 - 14x_2 + 6x_1x_2 + 3x_2^2\right)\right] \times$ $\left[30 + \left(2x_1 - 3x_2\right)^2 \times\right.$ $\left.\left(18 - 32x_1 + 12x_1^2 + 48x_2 - 36x_1x_2 + 27x_2^2\right)\right]$	$-2 \le x_i \le 2$	3
$f_{19}(x) = -\sum_{i=1}^{4}\exp\left[-\sum_{j=1}^{n} a_{ij}\left(x_j - p_{ij}\right)^2\right]$	$0 \le x_i \le 1$	-3.86
$f_{20}(x) = -\sum_{i=1}^{4}\exp\left[-\sum_{j=1}^{n} a_{ij}\left(x_j - p_{ij}\right)^2\right]$	$0 \le x_i \le 1$	-3.32

In Figures 14A, B, the $-2N$ force and $-1N.m$ torque are applied to the robot. So the control inputs are $2N$ and $1N.m$ to counteract the effect of the external force and torque on the robot.

The test functions for swarm intelligence optimization algorithms are shown in Tables 5–7.

5. Conclusion

This paper proposes an RBFNN-based anti-input saturation AFTC to solve the problem of degraded control performance of the CDR during movement on the water surface caused by drive faults, uncertain water resistance, and uncertain model parameters. The AFTC incorporates a fast NTSMC, which ensures the robustness of the robot against external disturbances and the effects of uncertain model parameters. The RBFNN is used to estimate drive faults and compensate for the controller output. Additionally, an anti-input saturation control algorithm is introduced to prevent controller input saturation. Furthermore, the traditional approach of manually tuning controller parameters based on the designer's experience and iterative debugging is replaced with an optimization method called HDSA. The HDSA algorithm optimizes the controller parameters to ensure the optimal control performance of the robot.

In further work, adaptive algorithms are necessary for the adjustment of the upper limit of the maximum control input to the robot on the ground and on the water surface.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

KW implementation and execution of the theory research and experiment and writing of the manuscript. YL theoretical support on the idea and helped write the manuscript. CH preliminary work and revising the manuscript. All authors actively contributed to the preparation of the content of this paper.

Funding

This work was supported in part by the Sharing Technology Project (41412040102), the China National Science Foundation (61473155), the Jiangsu Technology Department under Modern Agriculture (BE2017301), and the Six Talent Peaks Project in Jiangsu Province (GDZB-039).

Acknowledgments

We thank the editors and reviewers of the journal.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ali, N., Tawiah, I., and Zhang, W. (2020). Finite-time extended state observer based nonsingular fast terminal sliding mode control of autonomous underwater vehicles. *Ocean Eng.* 218, 108179. doi: 10.1016/j.oceaneng.2020.108179

Chen, G., Tu, J., Ti, X., Wang, Z., and Hu, H. (2021). Hydrodynamic model of the beaver-like bendable webbed foot and paddling characteristics under different flow velocities. *Ocean Eng.* 234, 109179. doi: 10.1016/j.oceaneng.2021.109179

Chen, L., Cui, R., Yang, C., and Yan, W. (2019). Adaptive neural network control of underactuated surface vessels with guaranteed transient performance: theory and experimental results. *IEEE Transact. Ind. Electron.* 67, 4024–4035. doi: 10.1109/TIE.2019.2914631

Chu, R., Liu, Z., and Chu, Z. (2022). Improved super-twisting sliding mode control for ship heading with sideslip angle compensation. *Ocean Eng.* 260, 111996. doi: 10.1016/j.oceaneng.2022.111996

Cohen, A., and Zarrouk, D. (2020). "The amphistar high speed amphibious sprawl tuned robot: design and experiments," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Las Vegas, NV: IEEE), 6411–6418.

Deng, Y., Zhang, X., Im, N., Zhang, G., and Zhang, Q. (2020). Adaptive fuzzy tracking control for underactuated surface vessels with unmodeled dynamics and input saturation. *ISA Trans.* 103, 52–62. doi: 10.1016/j.isatra.2020.04.010

Dorigo, M., Maniezzo, V., and Colorni, A. (1996). Ant system: optimization by a colony of cooperating agents. *IEEE Transact. Syst. Man Cybernet. Part B* 26, 29–41. doi: 10.1109/3477.484436

Fister, I., Fister Jr, I., Yang, X.-S., and Brest, J. (2013). A comprehensive review of firefly algorithms. *Swarm Evol. Comput.* 13, 34–46. doi: 10.1016/j.swevo.2013.06.001

Gao, B., Liu, Y.-J., and Liu, L. (2022). Adaptive neural fault-tolerant control of a quadrotor uav via fast terminal sliding mode. *Aerospace Sci. Technol.* 129, 107818. doi: 10.1016/j.ast.2022.107818

Gheisarnejad, M., and Khooban, M. H. (2020). An intelligent non-integer pid controller-based deep reinforcement learning: Implementation and experimental results. *IEEE Transact. Ind. Electron.* 68, 3609–3618. doi: 10.1109/TIE.2020.2979561

Guo, J., Zhang, K., Guo, S., Li, C., and Yang, X. (2019). "Design of a new type of tri-habitat robot," in *2019 IEEE International Conference on Mechatronics and Automation (ICMA)* (Tianjin: IEEE), 1508–1513.

Guo, X., Huang, S., Lu, K., Peng, Y., Wang, H., and Yang, J. (2022). A fast sliding mode speed controller for PMSM based on new compound reaching law with improved sliding mode observer. *IEEE Trans. Transp. Elect.* 9, 2955–2968.

Heidari, A. A., Mirjalili, S., Faris, H., Aljarah, I., Mafarja, M., and Chen, H. (2019). Harris hawks optimization: Algorithm and applications. *Fut. Gen. Comp. Syst.* 97, 849–872. doi: 10.1016/j.future.2019.02.028

Hou, Q., and Ding, S. (2021). Finite-time extended state observer-based super-twisting sliding mode controller for pmsm drives with inertia identification. *IEEE Transact. Transport. Electrif.* 8, 1918–1929. doi: 10.1109/TTE.2021.3123646

Huang, J., Wang, W., Wen, C., and Li, G. (2019). Adaptive event-triggered control of nonlinear systems with controller and parameter estimator triggering. *IEEE Trans. Automat. Contr.* 65, 318–324. doi: 10.1109/TAC.2019.2912517

Jiang, T., and Lin, D. (2020). Fast finite-time backstepping for helicopters under input constraints and perturbations. *Int. J. Syst. Sci.* 51, 2868–2882. doi: 10.1080/00207721.2020.1803438

Liao, Y.-,l., Zhang, M.-,j., Wan, L., and Li, Y. (2016). Trajectory tracking control for underactuated unmanned surface vehicles with dynamic uncertainties. *J. Cent. South Univ.* 23, 370–378. doi: 10.1007/s11771-016-3082-4

Liu, K., Gao, H., Ji, H., and Hao, Z. (2020). Adaptive sliding mode based disturbance attenuation tracking control for wheeled mobile robots. *Int. J. Control Automat. Syst.* 18, 1288–1298. doi: 10.1007/s12555-019-0262-7

Liu, X., Zhang, M., and Yao, F. (2018). Adaptive fault tolerant control and thruster fault reconstruction for autonomous underwater vehicle. *Ocean Eng.* 155, 10–23. doi: 10.1016/j.oceaneng.2018.02.007

Mirjalili, S. (2016). Sca: a sine cosine algorithm for solving optimization problems. *Knowl. Based Syst.* 96, 120–133. doi: 10.1016/j.knosys.2015.12.022

Mirjalili, S., Mirjalili, S. M., and Lewis, A. (2014). Grey wolf optimizer. *Adv. Eng. Softw.* 69, 46–61. doi: 10.1016/j.advengsoft.2013.12.007

Najafi, A., Vu, M. T., Mobayen, S., Asad, J. H., and Fekih, A. (2022). Adaptive barrier fast terminal sliding mode actuator fault tolerant control approach for quadrotor uavs. *Mathematics* 10, 3009. doi: 10.3390/math10163009

Nan, F., Sun, S., Foehn, P., and Scaramuzza, D. (2022). Nonlinear mpc for quadrotor fault-tolerant control. *IEEE Robot. Automat. Lett.* 7, 5047–5054. doi: 10.1109/LRA.2022.3154033

Shen, Q., Yue, C., Goh, C. H., and Wang, D. (2018). Active fault-tolerant control system design for spacecraft attitude maneuvers with actuator saturation and faults. *IEEE Transact. Ind. Electron.* 66, 3763–3772. doi: 10.1109/TIE.2018.2854602

Song, M.-P., and Gu, G.-C. (2004). "Research on particle swarm optimization: a review," in *Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No. 04EX826)*, (Shanghai: IEEE), 2236–2241.

Wang, F., Ma, Z., Gao, H., Zhou, C., and Hua, C. (2022). Disturbance observer-based nonsingular fast terminal sliding mode fault tolerant control of a quadrotor UAV with external disturbances and actuator faults. *Int. J. Cont. Autom. Syst.* 20, 1122–1130. doi: 10.1007/s12555-020-0773-2

Wang, H., Shi, J., Wang, J., Wang, H., Feng, Y., and You, Y. (2019a). Design and modeling of a novel transformable land/air robot. *Int. J. Aero. Eng.* doi: 10.1155/2019/2064131

Wang, K., Liu, Y., Huang, C., and Bao, W. (2022a). Water surface flight control of a cross domain robot based on an adaptive and robust sliding mode barrier control algorithm. *Aerospace* 9, 332. doi: 10.3390/aerospace9070332

Wang, K., Liu, Y., Huang, C., and Cheng, P. (2022b). Water surface and ground control of a small cross-domain robot based on fast line-of-sight algorithm and adaptive sliding mode integral barrier control. *Appl. Sci.* 12, 5935. doi: 10.3390/app12125935

Wang, N., and Deng, Z. (2019). Finite-time fault estimator based fault-tolerance control for a surface vehicle with input saturations. *IEEE Trans. Ind. Informat.* 16, 1172–1181. doi: 10.1109/TII.2019.2930471

Wang, N., Xie, G., Pan, X., and Su, S.-F. (2019b). Full-state regulation control of asymmetric underactuated surface vehicles. *IEEE Trans. Ind. Informat.* 66, 8741–8750. doi: 10.1109/TIE.2018.2890500

Wu, G., Chen, G., Zhang, H., and Huang, C. (2021). Fully distributed event-triggered vehicular platooning with actuator uncertainties. *IEEE Transact. Vehic. Technol.* 70, 6601–6612. doi: 10.1109/TVT.2021.3086824

Wu, L.-B., Park, J. H., Xie, X.-P., Gao, C., and Zhao, N.-N. (2020). Fuzzy adaptive event-triggered control for a class of uncertain nonaffine nonlinear systems with full state constraints. *IEEE Transact. Fuzzy Syst.* 29, 904–916. doi: 10.1109/TFUZZ.2020.2966185

Xing, H., Shi, L., Hou, X., Liu, Y., Hu, Y., Xia, D., et al. (2021). Design, modeling and control of a miniature bio-inspired amphibious spherical robot. *Mechatronics* 77, 102574. doi: 10.1016/j.mechatronics.2021.102574

Xue, J., and Shen, B. (2020). A novel swarm intelligence optimization approach: sparrow search algorithm. *Syst. Sci. Control Eng.* 8, 22–34. doi: 10.1080/21642583.2019.1708830

Xue, J., and Shen, B. (2022). Dung beetle optimizer: a new meta-heuristic algorithm for global optimization. *J. Supercomput.* 1–32. doi: 10.1007/s11227-022-04959-6

Yu, X.-N., Hao, L.-Y., and Wang, X.-L. (2022). Fault tolerant control for an unmanned surface vessel based on integral sliding mode state feedback control. *Int. J. Control Automat. Syst.* 20, 2514–2522. doi: 10.1007/s12555-021-0526-x

Zhang, G., Chu, S., Zhang, W., and Liu, C. (2022). Adaptive neural fault-tolerant control for usv with the output-based triggering approach. *IEEE Transact. Vehic. Technol.* 71, 6948–6957. doi: 10.1109/TVT.2022.3167038

Zhang, H., Xi, R., Wang, Y., Sun, S., and Sun, J. (2021). Event-triggered adaptive tracking control for random systems with coexisting parametric uncertainties and severe nonlinearities. *IEEE Trans. Automat. Contr.* 67, 2011–2018. doi: 10.1109/TAC.2021.3079279

Zhao, Y., Qi, X., Ma, Y., Li, Z., Malekian, R., and Sotelo, M. A. (2020). Path following optimization for an underactuated usv using smoothly-convergent deep reinforcement learning. *IEEE Transact. Intell. Transport. Syst.* 22, 6208–6220. doi: 10.1109/TITS.2020.2989352

Zhong, G., Cao, J., Chai, X., and Bai, Y. (2021). Design and performance analysis of a triphibious robot with tilting-rotor structure. *IEEE Access* 9, 10871–10879. doi: 10.1109/ACCESS.2021.3050182

frontiers | Frontiers in Neurorobotics

TYPE Original Research
PUBLISHED 02 August 2023
DOI 10.3389/fnbot.2023.1193823

Check for updates

OPEN ACCESS

EDITED BY
Long Jin,
Lanzhou University, China

REVIEWED BY
Ning Sun,
Nankai University, China
Bo Yang,
University of Electronic Science and Technology of China, China
Bolin Liao,
Jishou University, China

*CORRESPONDENCE
Jianwen Guo
✉ guojw@dgut.edu.cn

RECEIVED 25 March 2023
ACCEPTED 17 July 2023
PUBLISHED 02 August 2023

CITATION
Xie L, Huang J, Li Y and Guo J (2023) An improved model for target detection and pose estimation of a teleoperation power manipulator. *Front. Neurorobot.* 17:1193823. doi: 10.3389/fnbot.2023.1193823

An improved model for target detection and pose estimation of a teleoperation power manipulator

Li Xie[1], Jiale Huang[1,2], Yutian Li[1] and Jianwen Guo[1]*

[1]School of Mechanical Engineering, Dongguan University of Technology, Dongguan, China, [2]School of Electromechanical Engineering, Guangdong University of Technology, Guangzhou, China

Introduction: A hot cell is generally deployed with a teleoperation power manipulator to complete tests, operations, and maintenance. The position and pose of the manipulator are mostly acquired through radiation-resistant video cameras arranged in the hot cell. In this paper, deep learning-based target detection technology is used to establish an experimental platform to test the methods for target detection and pose estimation of teleoperation power manipulators using two cameras.

Methods: In view of the fact that a complex environment affects the precision of manipulator pose estimation, the dilated-fully convolutional one-stage object detection (dilated-FCOS) teleoperation power manipulator target detection algorithm is proposed based on the scale of the teleoperation power manipulator. Model pruning is used to improve the real-time performance of the dilated-FCOS teleoperation power manipulator target detection model. To improve the detection speed for the key points of the teleoperation power manipulator, the keypoint detection precision and model inference speed of different lightweight backbone networks were tested based on the SimpleBaseline algorithm. MobileNetv1 was selected as the backbone network to perform channel compression and pose distillation on the upsampling module so as to further optimize the inference speed of the model.

Results and discussion: Compared with the original model, the proposed model was experimentally proven to reach basically the same precision within a shorter inference time (only 58% of that of the original model). The experimental results show that the compressed model basically retains the precision of the original model and that its inference time is 48% of that of the original model.

KEYWORDS

teleoperation power manipulator, camera, target detection, pose estimation, deep learning

1. Introduction

Hot cells in nuclear power plants and high-energy physics devices are shielded from radiation (Zheng et al., 2015; Zhang et al., 2022), and they play a crucial role in testing, operation, and maintenance activities. To facilitate tasks such as inspection, assembly, disassembly, transportation, and part repair, hot cells are equipped with either a master-slave manipulator or a teleoperation power manipulator (Pezhman and Saeed, 2011; Assem et al., 2014; Zhang et al., 2021). These manipulators are necessary to mitigate the harmful effects of radiation on humans. To assist the teleoperator, the teleoperation power manipulator relies on sensing technologies, including visual sensing (Maruyama et al., 2014) and force sensing (Oosterhout et al., 2012), to gather information about the operation area.

In hot cells, where the radiation environment limits the use of certain sensors, radiation-resistant cameras are commonly installed to capture on-site images and transmit them to operators via the network. To regularly replace single modules in a tokamak vessel, Qiu et al. (2016) used a hand–eye coordination method to ensure the consistency between the operator's hand movement and the manipulator's end effector movement. Ribeiro et al. (2020) designed a hand–eye camera system for the acquisition of key information in the operating environment. Lionel et al. (2018) introduced the virtual reality technology in the assembly and tooling design of the tokamak diverter to assist teleoperators and successfully achieved assembly with a gap of <1 mm. Ferreira et al. (2012) designed a localization system based on cameras to accurately estimate the position and direction of CPRH by capturing video streams for the implementation of an augmented reality system. Liu et al. (2020) proposed the vision-based breakpoint detection algorithm and successfully identified and captured tiles that had fallen onto the diverter by employing the watershed segmentation algorithm.

Most of the information about the teleoperation power manipulator's position and pose comes from radiation-resistant cameras in the hot cell. The operator's teleoperation efficiency is impacted by the limited visual information provided by this method of observation solely by human eyes through cameras. The application of technologies such as virtual reality (VR) or augmented reality (AR) can integrate the information of cameras into the operation platform of VR or AR, which is conducive to improving the operation efficiency (Qiu et al., 2016; Lionel et al., 2018; Ribeiro et al., 2020). However, obtaining the teleoperation power manipulator's position and pose from the photographs is one of the issues that need to be resolved in the hot cell.

The deep learning-based pose estimation algorithm can quickly distinguish poses from RGB images and achieve satisfactory estimation results. Kehl et al. (2017) proposed a direct regression-based 6D pose estimation method to achieve end-to-end 6D pose estimation. DeePose (Toshev and Szegedy, 2014) applied a convolutional neural network (CNN) to human pose estimation for the first time and achieved higher precision than traditional methods. Pose coordinate regression-based algorithms, on the other hand, only constrain the pose coordinates with the mean square error and ignore the supervision of the spatial information of the key points, making it difficult to further improve their regression precision. Wei et al. (2016) proposed a sequential architecture composed of convolutional networks to predict the locations of the key points and introduced the key points heatmap as the input of the next stage, which provides rich spatial information for the subsequent network layer and improves the robustness of the algorithm. Sun et al. (2017) proposed HRNet, which is composed of multi-resolution subnetworks connected in parallel and achieved the best pose estimation results on the COCO dataset in 2019. Mišeikis et al. (2018a,b) proposed a multi-objective CNN, which uses 2D images to estimate the 3D positions of the key points and used transfer learning techniques to adapt the CNN trained to estimate the poses of UR robots to Kuka robots. Heindl et al. (2019) proposed a multi-robot pose estimation method based on a recurrent neural network, which uses 2D images as input and simultaneously infers the number of robots in the scene, the joint locations, and the sparse depth maps around the joint locations, demonstrating high generalizability to the real-world images. Ning et al. (2020) presents a real-time 3D face-alignment method that uses an encoder-decoder network with an efficient deconvolution layer which has low prediction errors with real-time applicability. Wu et al. (2022) presents an age-compensated makeup transformation framework based on homology continuity, and the experimental results show that the framework outperforms existing methods.

The technical conditions for the pose estimation of teleoperation power manipulators are provided by the aforementioned studies. In this paper, target detection and pose estimation of teleoperation power manipulators are designed based on deep learning, obtaining the teleoperation power manipulator's position and pose by two cameras in the hot cell, which is few studied in this field at present. A dilated-fully convolutional one-stage object detection (dilated-FCOS) target detection algorithm for teleoperation power manipulators is suggested in accordance with its scale. For teleoperation power manipulators, a keypoint detection algorithm based on SimpleBaseline has been developed. This algorithm reduces the model's inference time while maintaining model precision. Through teleoperation power manipulator pose estimation experiments, an experimental platform for teleoperation power manipulator operation is established to confirm the methods' viability and efficacy.

The following is the layout of the remainder of the paper: the construction of the experimental platform and the production of the experimental data are both covered in detail in Section 2; the proposed dilated-FCOS teleoperation power manipulator target detection method is presented in Sections 3; the keypoint detection method is in the Section 4; experiments and discussion are the main focus of Section 5; summary of this work and suggestions for future research are presented in Section 6.

2. Experimental platform and experimental data

2.1. Construction of the experimental platform

The experimental platform (Figure 1) consists of several components: a teleoperation power manipulator, a camera system with two cameras, a motion capture system, an image processing module, and a teleoperation power manipulator display module. The camera system captures real-time operational images of the teleoperation power manipulator, while the image processing module detects targets and estimates the pose of the manipulator. The updated pose information is then inputted into the teleoperation power manipulator display module to adjust its position accordingly.

(1) Teleoperation power manipulator. Figure 2C depicts the teleoperation power manipulator for teleoperation. It is configured with eight degrees of freedom, consisting of four rotational and four translational degrees of freedom. The mobile platform, depicted in Figure 2A in two dimensions, allows the teleoperation power manipulator to move forwards and backwards to reach the desired operational position. Figure 2B presents the 3D model of the

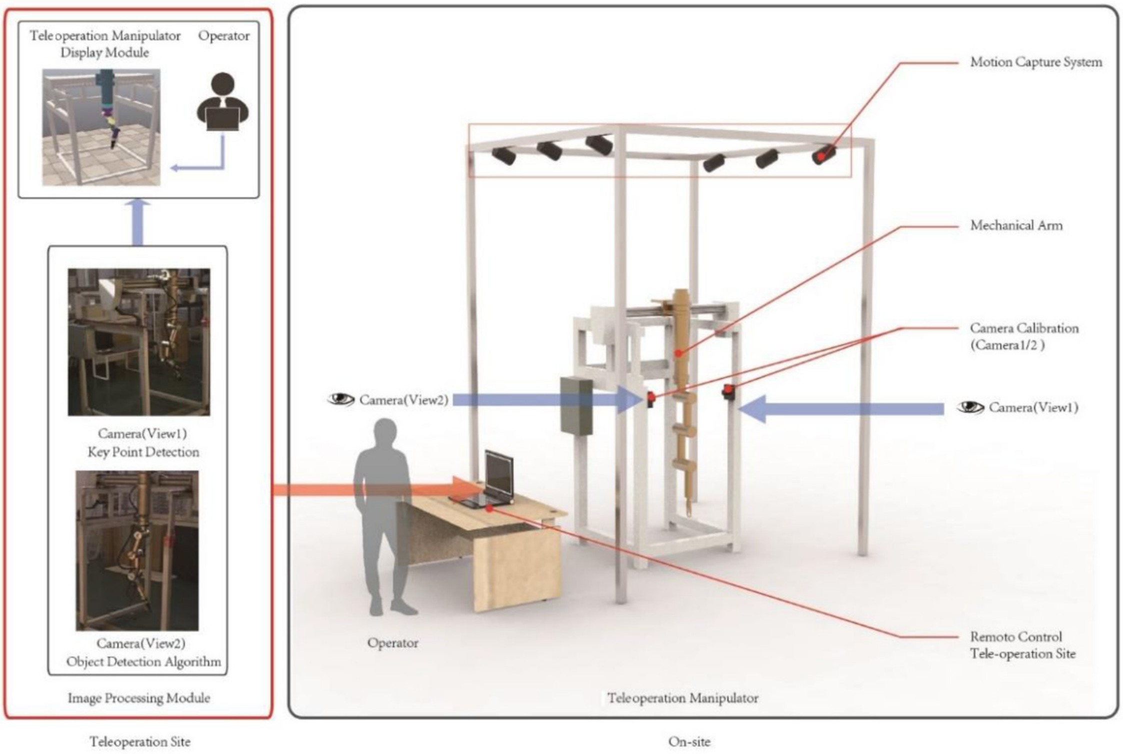

Teleoperation Manipulator Display Module
Operator
Camera(View1) Key Point Detection
Camera(View2) Object Detection Algorithm
Image Processing Module
Teleoperation Site
Motion Capture System
Mechanical Arm
Camera Calibration (Camera1/2)
Camera(View2)
Camera(View1)
Operator
Remoto Control Tele-operation Site
Teleoperation Manipulator
On-site

FIGURE 1
Experimental platform.

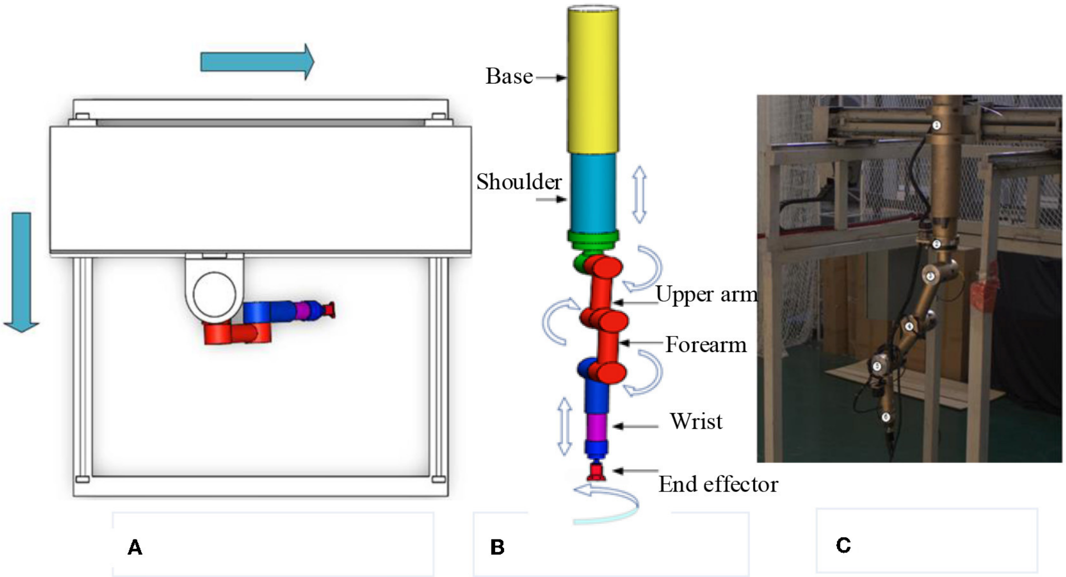


FIGURE 2
Model of the teleoperation power manipulator. **(A)** Mobile platform model. **(B)** Manipulator model. **(C)** Real manipulator.

teleoperation power manipulator, which includes a base, a shoulder, an upper arm, a forearm, a wrist, and an end effector.

(2) Motion capture system. Camera calibration and the creation of a global coordinate system that is parallel to the mobile platform's translational direction are both made easier by the motion capture system. The OptiTrack system (Motive Optical motion capture software., 2023) is the motion capture system used in this paper.

TABLE 1 OPT-CC500-GM-04 camera parameters.

Parameter type	Parameter value
Data interface	GigE
Resolution	2,448(H)*2,048(V)
Chip size	2/3"
Maximum frame rate	30 fps
Pixel size	3.45 µm
Exposure time	34 µs-1 s
Optical interface	C
Size	29 mm x 29 mm x 42 mm

(3) Camera system. The camera system consists of two industrial cameras, which capture the operational status of the teleoperation power manipulator from two different angles. Table 1 provides the specific parameters of the cameras, including a focal length of 16 mm, a distortion rate of <0.2%, and a resolution of 5 million pixels.

(4) Image processing module. The function of the image processing module is to locate the teleoperation power manipulator in a complex environment through the target detection algorithm, send the relevant information to the key points detection network for pose estimation, and input the pose information into the teleoperation power manipulator display module. The angles of the rotation joints of the teleoperation power manipulator are calculated based on the angles between the vectors formed by every two key points (O'Donovan et al., 2006). The translational joints are located by determining the translational distances of the key points in the 3D space through multiview-based triangulation (Zeng et al., 1999). The target detection and pose estimation methods of this module are the main research contents of this paper.

(5) The teleoperation power manipulator display module. The module was developed using Python and the V-REP Robot Simulator (Liu et al., 2017). To accurately represent the real teleoperation power manipulator, a model was created in Solidworks and subsequently imported into V-REP. The multiview teleoperation power manipulator pose estimation model is then utilized to continuously update the virtual teleoperation power manipulator's translational distances and pose information.

2.2. Preparation of the training dataset

To build a teleoperation power manipulator target detection model, the sample data for training the target detection model must be prepared first. The sample data are prepared in the following two steps:

(1) Acquisition of moving images of the teleoperation power manipulator

The image data are acquired mainly through the continuous acquisition of moving images of the teleoperation power manipulator from different angles through two cameras. To improve the robustness of the model, data were collected under different lighting conditions.

(2) Dataset labeling

The key points (namely, the base, the shoulder, the upper arm, the forearm, and the wrist) of the teleoperation power manipulator are shown in Figure 3. The labeling tool LabelImg and the Visual Object Classes (VOC) Format are utilized in this paper. With reference to the MPII human pose estimation dataset (Simon et al., 2016), the files are labeled with the visibility and coordinates of the five key points. In addition, to improve the ability of the model to detect occluded key points, the slightly occluded key points were labeled and set to be visible. The different positions of the teleoperation power manipulator have different degrees of illumination during the operation. Color dithering is used to boost the robustness of the model to illumination, and random noise is added to the data to boost the model's robustness. The total number of samples generated was 4,000. The numbers of samples in the training set and the test set obtained after random allocation of the total samples were 3,600 and 400, respectively.

3. Dilated-FCOS method

Fully Convolutional One-Stage Object Detection (FCOS) (Coppelia Robotics GmbH, 2022) is a fully convolutional anchor-free single-stage target detection algorithm. To suit the application of teleoperation power manipulator, a dilated-FCOS teleoperation power manipulator target detection method, is proposed. The structure of dilated-FCOS is shown in Figure 4.

(1) The improved network structure of the FCOS. According to the characteristics of the large target in teleoperation power manipulator detection, the FCOS network structure is modified to improve the detection precision, to reduce the time required for feature extraction, and to increase the model inference speed.

(2) Channel pruning of the FCOS. The FCOS target detection model's backbone network (darknet19) was optimized with the channel pruning algorithm to make it more precise and effective due to its high parameter redundancy and high computational overhead.

3.1. Method

3.1.1. FCOS network

The structure of the FCOS network is shown in Figure 5. Darknet-19, the backbone network of FCOS (Andriluka et al., 2014), outputs three scale outputs (C3, C4, C5), and the feature pyramid outputs five scale outputs (P3, P4, P5, P6, P7). P3 is a high-resolution feature map with rich spatial information. P4 focuses on the detection of small targets. P5, P6, and P7 are low-resolution feature maps with rich semantic information, which focus on the detection of large and medium targets. The design concept of FCOS is divided into the following points:

(1) Pixel by pixel for the detection. Anchor-based algorithms often rely on artificially designing a significant number of anchor frames to enhance the recall rate. However, this approach introduces a challenge of imbalance between positive and negative samples during training, as the majority of anchor frames are negative samples. Additionally, the calculation complexity increases due to the intersection ratio between all anchor frames

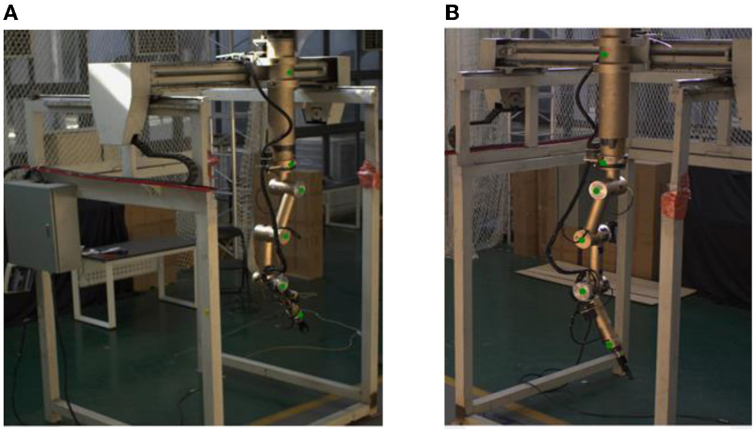


FIGURE 3
Keypoint labeling. **(A)** Camera(View)1. **(B)** Camera(View)2.

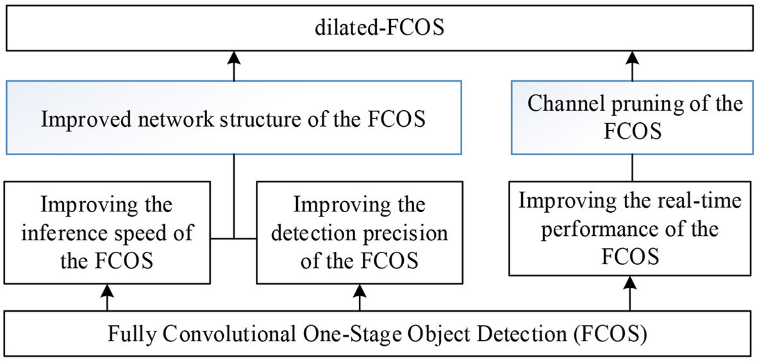

dilated-FCOS
Improved network structure of the FCOS
Channel pruning of the FCOS
Improving the inference speed of the FCOS
Improving the detection precision of the FCOS
Improving the real-time performance of the FCOS
Fully Convolutional One-Stage Object Detection (FCOS)

FIGURE 4
Dilated-FCOS method.

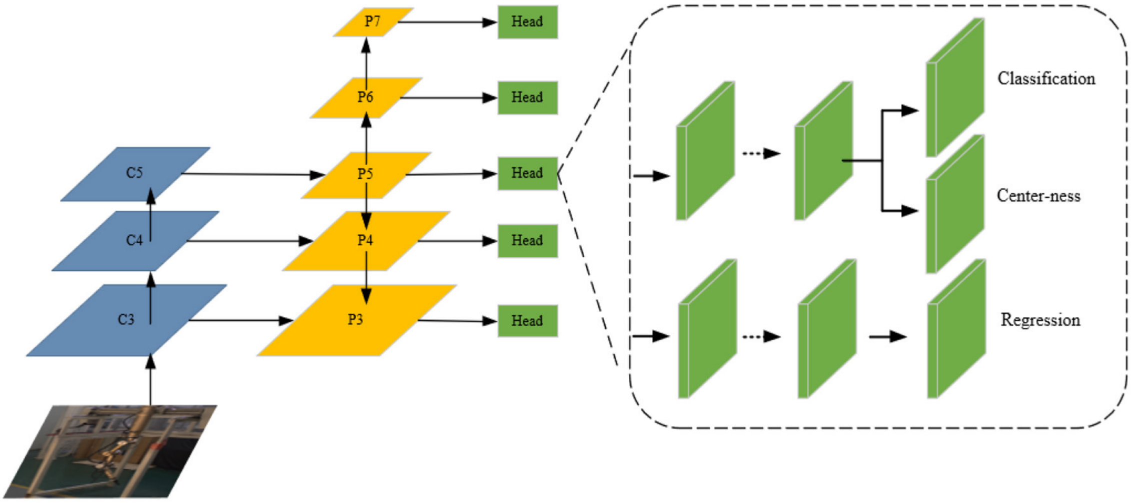


FIGURE 5
Network structure of the FCOS.

and boundary boxes during training. In contrast, FCOS is an anchor-free algorithm that avoids the use of anchor frames. Instead, it maps each feature point on the feature map to the original map and performs regression. By incorporating a larger number of positive samples, FCOS facilitates improved model training and leads to significant enhancements in the detector's performance.

(2) Multi-scale training strategy

Deep network has rich semantic information, that is, the output result is not affected by the position of the feature graph, which is suitable for classification task; the shallow feature has rich spatial information, that is, the output result changes according to the change of the features, which is suitable for regression task. Target detection requires both the regression of the target location and the target classification. To solve these two contradictory tasks simultaneously, FCOS adopts a feature pyramid structure to fuse the feature maps at different scales, so that the semantic information and spatial information between the different feature maps can complement each other. The feature pyramid network structure is shown in the Figure 6. The first part of the network is the path from the bottom, the backbone network, and the path is the lack of spatial information, and the features, adding the spatial information and semantic information of the feature map. In the third part, the lateral connection path adjusts the number of channels in the fusion to perform prediction and regression tasks. Integrating the information of different scales, the feature pyramid greatly improves the target detection accuracy of FCOS.

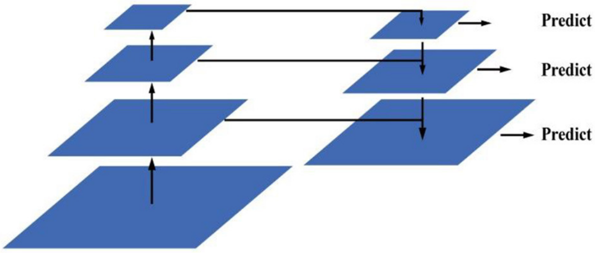


FIGURE 6
Feature pyramid networks.

(3) Center confidence degree prediction

As shown in Figure 6, the central confidence degree is a branch increased in the prediction of each test head. The calculation of the central confidence is such as formulas (1). The detection box away from the central point is optimized by the cross entropy loss function. By combining the boundary box away from the object with the non-maximum suppression, the detection performance is significantly improved.

$$centerness = \sqrt{\frac{\min(l^*, r^*)}{\max(l^*, r^*)} \times \frac{\min(t^*, b^*)}{\max(t^*, b^*)}} \tag{1}$$

where l^*, r^*, and b^* are the distance from the sampling point to the four sides of the boundary box.

3.1.2. The improved network structure of the FCOS

Figure 7 shows the improved network structure of the FCOS with two major improvements.

(1) Improving the detection precision of the FCOS. Teleoperation power manipulator detection is a form of large target detection. Considering that the C5 feature layer has a limited detection scale range, a dilated encoder (Tian et al., 2020)

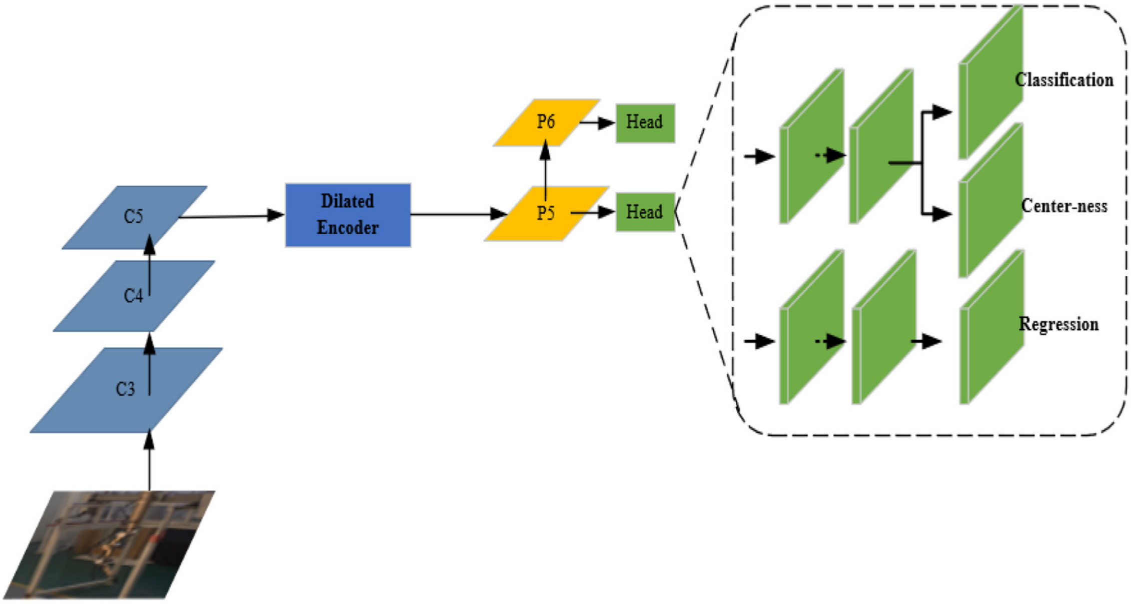


FIGURE 7
The improved network structure of the FCOS.

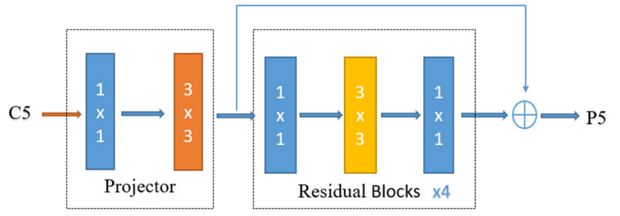


FIGURE 8
Dilated encoder.

is introduced to enrich the receptive field of the C5 feature layer by stacking continuous dilated residual units, and P6 is retained to improve the robustness of large target detection. The dilated encoder is shown in Figure 8. The first part of the encoder reduces the number of output channels through a 1 × 1 convolutional layer and then extracts semantic feature information through a 3 × 3 convolutional layer. The second part enlarges the receptive field through stacking continuous 3 × 3 dilated residual units with different dilation rates.

(2) Improving the inference speed of the model. Detecting the shallow features of small targets has very little effect on large targets such as teleoperation power manipulators. The shallow feature maps (P3, P4) are discarded here to improve the detection speed of the FCOS, and the P7 feature layer is discarded to improve the real-time performance of the network model. The improved network only performs the final classification, position regression, and central confidence interval prediction on the feature maps P5 and P6.

3.1.3. Channel pruning of the FCOS

Channel pruning (Redmon and Farhadi, 2017) is a method that improves the real-time performance of a model by compressing the model. Through sparse training on the channel scaling factor, channel pruning leads to channel sparsification.

Adding a batch normalization (BN) layer (Chen et al., 2021) after the convolutional layer can achieve rapid convergence and better generalization performance. The calculation formulas of BN are as follows:

$$\begin{aligned} \widehat{Z} &= \frac{Z_{in}-\mu}{\sqrt{\sigma^2+\varepsilon}} \\ Z_{out} &= \gamma\widehat{Z}+\beta \end{aligned} \tag{2}$$

where Z_{in} is the input tensor, Z_{out} is the output tensor, $\boldsymbol{\mu}$ is the vector of the mean value of the convolution result of each channel, σ is the vector of the variance of the convolution result, $\boldsymbol{\varepsilon}$ is a constant, $\boldsymbol{\gamma}$ is the learnable scaling factor in the BN layer, and $\boldsymbol{\beta}$ represents the learnable bias coefficient in the BN layer.

In Formula (2), when $\boldsymbol{\gamma}$ approaches 0, the effect of Z_{in} on Z_{out} is negligible. Here, $\boldsymbol{\gamma}$ is used as the scaling factor, and the parameter $\boldsymbol{\gamma}$ is penalized to save computational overhead and to avoid introducing unnecessary parameters.

The steps of channel pruning are as follows: (1) put all image data samples into the optimal model for sparse training; (2) sort the scaling factor $\boldsymbol{\gamma}$ of each BN layer; (3) prune the convolution layer corresponding to the scaling factor that has little effect on model performance; and (4) fine-tune the new model obtained by pruning to improve the detection performance of the network.

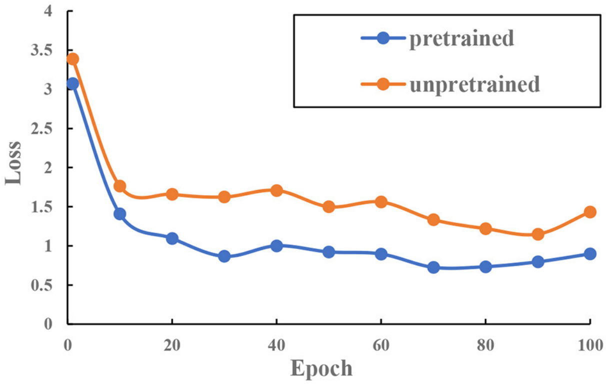


FIGURE 9
The loss–epoch curve.

3.2. Test of dilated-FCOS

3.2.1. Effectiveness test of the pre-trained model

Darknet19 is designed for ImageNet (Krizhevsky et al., 2012). Compared with the ImageNet dataset, the teleoperation power manipulator dataset is relatively small in size. Therefore, we first load and pre-train darknet19 on ImageNet to obtain the network weights to improve the network convergence speed. Two sets of experiments are set up to verify the effectiveness of the pre-trained model. Experiment 1 uses random weights to initialize the network, while Experiment 2 uses pre-trained weights on ImageNet to initialize the network. Both sets of experiments used the same learning strategy and optimization method. After 100 iterations, the loss curve was obtained, as shown in Figure 9. The results show that loading the pre-trained model can accelerate the model convergence.

3.2.2. Performance test of target detection

FCOS, Faster-RCNN (Ren et al., 2016), and dilated-FCOS were used for the target detection performance test. In the experiment, the mean average precision (mAP) (Henderson and Ferrari, 2017) was used to measure the target detection performance of the model, and the inference time (ms) was used to measure the inference speed of the model. The intersection over union (IoU) threshold was set to 0.5, and a uniform image input size of 640 × 640 was used in all three models. The test results shown in Table 2 indicate that the dilated-FCOS is superior to the FCOS in both model precision and inference time.

To further test the robustness of the network, two sets of experiments were conducted in this study. In the first set, 640 x 640 images with color perturbations were used as inputs, while in the second set, images with noise interference were fed into the network. The partial experimental results, as shown in Figure 10, demonstrate that the network exhibits excellent anti-interference ability.

3.2.3. Performance test of channel pruning

The first step of channel pruning is sparse training and screening out the channel numbers that have little impact on the output result. It is necessary to set the sparsity coefficient λ.

TABLE 2 Performance comparison of different models.

Model	mAP (%)	Inference time (ms)
FCOS	93.78	31.59
Faster-RCNN	96.38	63.52
Dilated-FCOS	95.24	23.86

Figure 11 shows the distribution of the scaling factor γ at different λ values. It can be seen that when $\lambda = 2$, γ is sparsified, but the effect is not obvious; when $\lambda = 5$, γ is close to 0, and the effect is obvious. Since $\lambda = 5$ is effective in screening the channel number, $\lambda = 5$ is selected to complete the sparse training.

Table 3 compares the performances of the model on the teleoperation power manipulator test set under different pruning rates. The original model has an mAP of 95.24%, a params of 35.96 M, and an inference time of 23.86 s on RTX 2080Ti. When the pruning rate is set to 0.1, the precision of the model increases slightly. This indicates that a higher precision can be achieved with fewer model parameters by removing the number of redundant channels of the original model. When the pruning rate is 0.1–0.6, the average precision of the model generally shows a slow downward trend. When the pruning rate is 0.6, the precision of the model reaches 92.78%. When the pruning rate is 0.7, the precision is reduced to

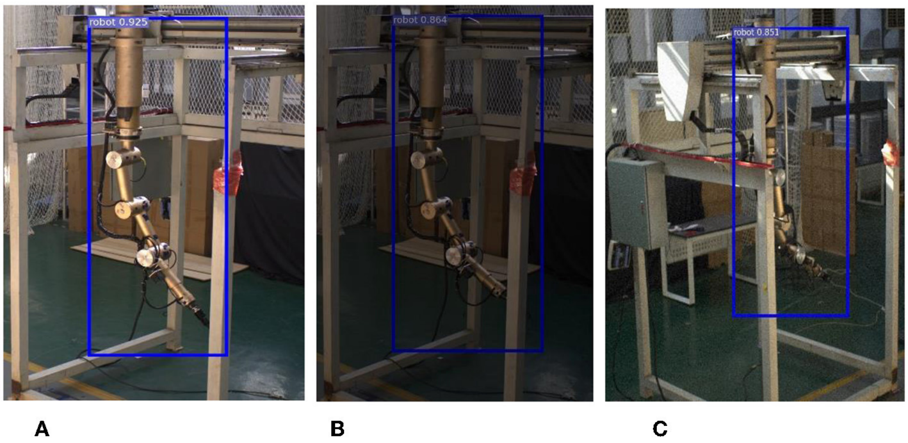


FIGURE 10
Results of network robustness. **(A)** Increase in brightness. **(B)** Decrease in brightness. **(C)** Adding noise.

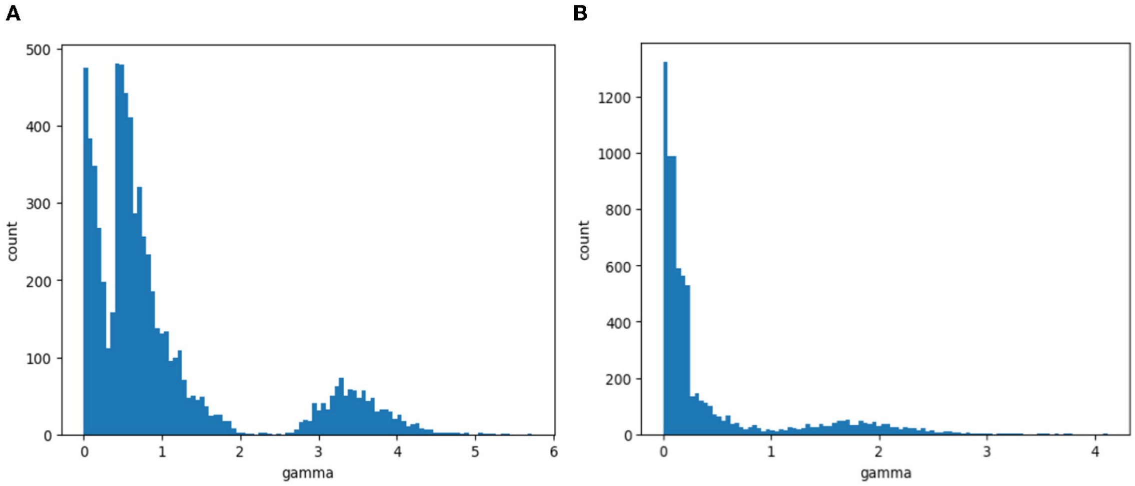


FIGURE 11
The distribution of γ at different λ values. **(A)** $\lambda = 2$. **(B)** $\lambda = 5$.

TABLE 3 The results of channel pruning.

Pruning ratio	mAP (%)	Params	Compression ratio	Inference time (Ms)
0	95.24	35.96 M	1	23.86
0.1	95.58	32.20 M	1.11	22.93
0.2	94.46	28.84 M	1.24	21.14
0.3	94.12	25.85 M	1.39	20.08
0.4	93.79	23.28 M	1.54	19.46
0.5	93.51	21.10 M	1.70	18.38
0.6	92.78	19.32 M	1.86	17.22
0.7	86.79	17.94 M	2.00	17.14

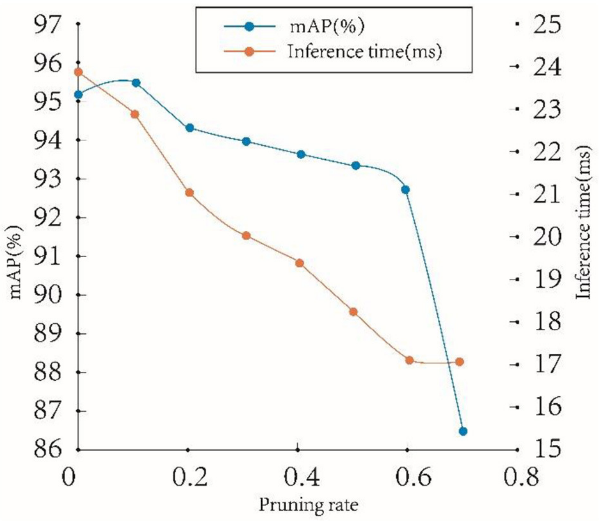


FIGURE 12
Different pruning rates of mAP and inference times.

86.79%. These results indicate that channel pruning can maintain the precision of the model within a certain range and will damage the precision of the model after exceeding a certain threshold.

Figure 12 shows the variation trend of the model precision and inference time on the teleoperation power manipulator dataset at different pruning rates. The model precision shows an upwards trend as the pruning rate increases from 0 to 0.1 and a gentle downward trend as the pruning rate increases from 0.3 to 0.6, while the inference time shows a more obvious downward trend as the pruning rate increases as the pruning rate increases, which indicates high model precision and small inference time delay at this time. When the pruning rate reaches 0.7, the precision decreases drastically, which indicates that pruning has severely damaged the precision of the model and has little effect on the optimization of the inference time. Therefore, the pruning rate is selected to be 0.5 in this paper to simultaneously achieve high precision and high inference speed.

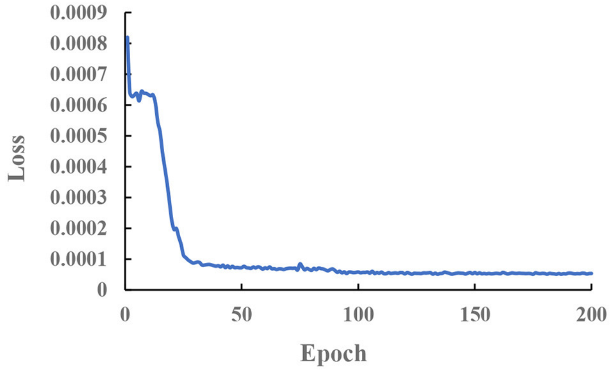


FIGURE 13
The diagram of loss.

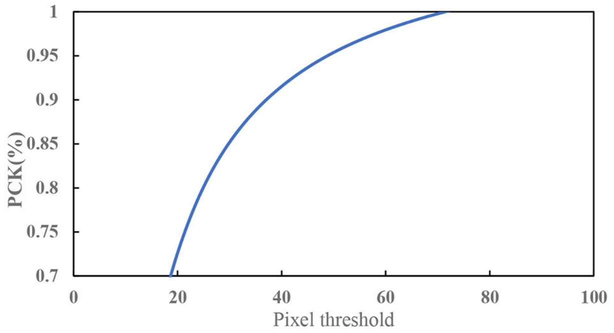


FIGURE 14
PCK under different pixel thresholds.

4. Keypoint detection method

SimpleBaseline (Xiao et al., 2018) is a simple and efficient 2D human keypoint detection network composed of the backbone network ResNet (Szegedy et al., 2016) and three transposed convolutions that are responsible for upsampling to restore the resolution. In this paper, a SimpleBaseline-lite-based keypoint detection method for teleoperation power manipulators is established through two main steps: replacing ResNet with a lightweight backbone network to improve the real-time performance of the model; compressing the channels of transposed convolutions to improve the inference speed of the model.

4.1. Setting of model training parameters

In this test, the PyTorch framework is used for model training, and the number of iterations is 140 epochs. The warmup strategy is used to improve the convergence speed of the model. The learning rate increases as the number of iterations increases and reaches the initial learning rate. The initial learning rate of the optimizer Adam is set to 0.001, and when the number of iterations reach 50 epochs, its learning rate decreases by 10-fold. The loss function is shown in Figure 13. The model can complete the convergence in 70 epochs.

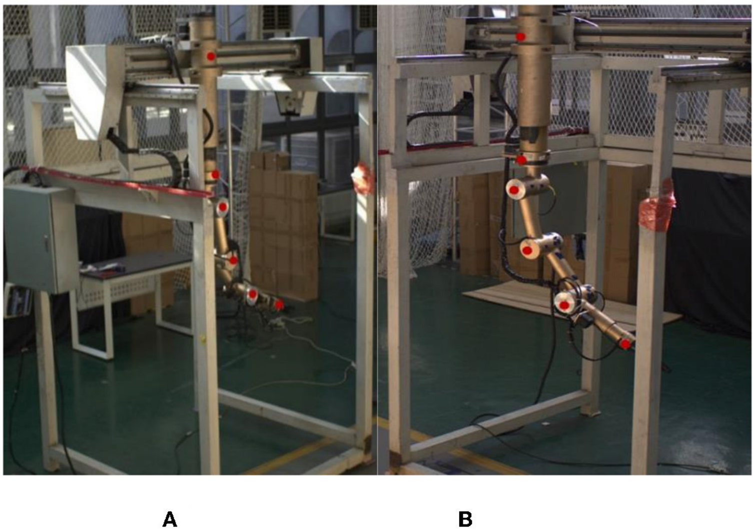


FIGURE 15
Visualization of prediction results on test sets. **(A)** Viewing angle 1. **(B)** Viewing angle 2.

In this paper, the Percentage of Correct key points (PCK) (Xiao et al., 2018) is used to analyse the detection performance of the SimpleBaseline network. PCK is the percentage of the predicted key points with a normalized distance from the ground truth that falls within the set threshold. PCK is calculated using formula (3).

$$\mathrm{PCK}{=}\frac{\sum_i \delta\left(\sqrt{(x_i-\widehat{x_i})+(y_i-\widehat{y_i})},\varepsilon\right)}{\sum_i 1} \qquad (3)$$
$$\delta\left(a,\varepsilon\right)=\left\{\frac{1,a\leq\varepsilon}{0,a\succ\varepsilon}\right.$$

where (x_i, y_i) are the 2D coordinates of a keypoint, $(\hat{x}_i, \hat{y}_i)$ are the 2D coordinates of the keypoint predicted by the network, and ε is the pixel threshold.

Figure 14 shows the PCK of the 2D key points of the teleoperation power manipulator under different pixel thresholds. The experimental results show that the PCK reaches 91.5% under the pixel threshold of 40. Figure 15 shows the distribution of the key points predicted by the network, which indicates that the SimpleBaseline network has a good detection effect on the key points of the teleoperation power manipulator.

4.2. Test and selection of lightweight convolutional networks

The lightweight feature networks MobileNetv1 (Howard et al., 2017), MobileNetv2 (Liu et al., 2018), MobileNetv3 (Howard et al., 2020), and ShuffleNetv2 (Ma et al., 2018) are used to replace ResNet50 as the feature extraction network and are tested on the teleoperation power manipulator dataset. The results are shown in Figure 16.

Figure 16 shows that among the four types of lightweight networks, the sparsity coefficient λ of MobileNetv2 and that of MobileNetv3 have a relatively large decrease. Based on the analysis of the network structure, MobileNetv2 has many depthwise separable convolutions compared with MobileNetv1 and introduces an inverted residual structure to solve the problem of the deactivation of depthwise separable convolutions. However, compared with the traditional convolution, the depthwise separable convolution extracts less effective feature information, resulting in the lack of spatial localization information and affecting the model precision.

The detection of the 2D key points of teleoperation power manipulators requires the contextual information of the feature map, which requires rich spatial information. For low-dimensional feature maps, the greater the number of channels is, the more abundant the spatial information. Resnet50, MobileNetv1, and ShuffleNetv2 have many channels in the low-dimensional network layer and can achieve good results in the detection of key points of teleoperation power manipulators.

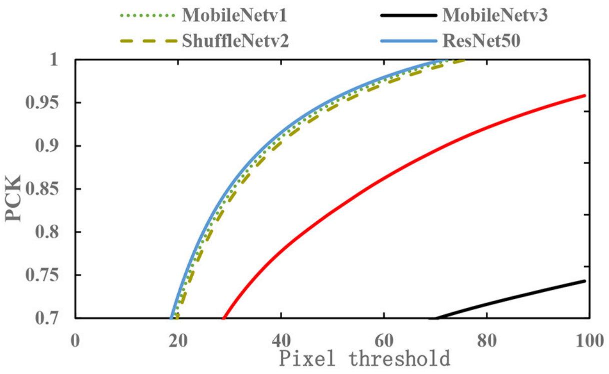


FIGURE 16
The PCK of different lightweight networks under different pixel thresholds.

Table 4 shows the test performances of different feature extraction networks. The input image size for both training and testing is 800 × 160. The params of MobileNetv1-SimpleBaseline is only 27% of the original value, the computational complexity is reduced to 55% of the original value, and the inference time is reduced to 56% of the original value. In summary, the MobileNetv1-SimpleBaseline network is selected in this paper.

4.3. Pose distillation

The upsampling module of SimpleBaseline is composed of three transposed convolutions with 256 channels. As the resolution of the upsampling feature map increases, the computational overheads of the transposed convolutions also increase. Compared with ResNet50, MobileNetv1 has an inferior feature extraction performance and sparser input features of the transposed convolutions. Keeping the number of channels in MobileNetv1 the same as that in ResNet may cause model redundancy and reduce the inference speed.

In this paper, the model is optimized by compressing the number of channels in the transposed convolutional layer. The number of channels of the three transposed convolutions is set to 64 n, 32 n, and 16 n, respectively, i.e., 384, 192, and 96 (n is the number of key points, which is set to 6). After compressing the number of channels, the computational complexity is reduced to 1/3 of the original value, and the params is reduced to 2/3 of the original value.

Table 5 compares the performance of the MobileNetv1-SimpleBaseline after compression of the number of channels (SimpleBaseline-a) with the performance of the uncompressed network. After channel compression, the model redundancy is reduced, and the parameters and computational complexity are greatly reduced. Although the computational overhead is greatly reduced, and the detection time is only 64% of that of the original model, the detection precision has reached 94% of that of the original model. This result shows that there are still redundant parameters in the upsampling module of SimpleBaseline-a. Based on this network, a model with higher precision is designed through pose distillation in this paper.

Pose distillation transfers the knowledge learned by a large network with good performance to a small network that is isomorphic or anti-isomorphic to the large network and compresses the model without significantly reducing the precision of the model (Hinton et al., 2014). The training process can be divided into two stages: training a powerful keypoint detection network as a teacher network and training a lightweight student model that simultaneously has high precision and high speed. The teacher model guides the student network to acquire high-level semantic information and strengthens the learning of the overall feature and spatial information by the student model.

Here, MobileNetv1-SimpleBaseline is selected as the teacher model, and SimpleBaseline-a is selected as the student model. The experiment is based on the PyTorch 1.5.1-GPU framework, the experimental operating system is Ubuntu 18.04, and the CUDA version is 10.2. The resolution of the network input image is 800 × 160, the initial learning rate is set to 0.001, the Adam optimizer is used, the batch size is set to 16, the momentum is set to 0.9, and the number of iterations is set to 140. The results are shown in Table 6.

Table 6 shows that the model precision of the student model after pose distillation was improved by 2%, but the parameters, computational complexity, and inference speed did not change. The results show that pose distillation can improve the detection precision of the key points of the teleoperation power manipulator.

The effectiveness of pose distillation is further illustrated by the visualized images in this paragraph. Figure 17 shows the predictions of the original student model (SimpleBaseline-a) and the student model after pose distillation (SimpleBaseline-lite) and the labeled visualized images. Occlusion and self-occlusion will inevitably occur in the teleoperation power manipulator (Figure 17A). Some occluded key points reduced the ability of the student model to extract spatial feature information, so the student model cannot fully learn the knowledge between channels of the feature map and the knowledge between the feature maps, resulting in a large deviation between the prediction result and the labels, which is the main reason for the decrease in detection precision. After the "tutoring" by the teacher model, as shown in Figure 17B,

TABLE 4 The performance of the lightweight SimpleBaseline in the test set.

Feature extraction network	Parameter	Computational complexity (GFLOPs)	Inference time (ms)	PCK@40 pixel(%)
ShuffleNetv2	7.54 M	12.97	22.73	90.3
MobileNetv1	9.50 M	14.07	18.17	90.9
MobileNetv2	9.56 M	13.92	21.2	83.4
MobileNetv3	5.57 M	11.37	23.93	81.1
Resnet50	33.99 M	25.18	30.21	91.5

TABLE 5 Comparison of the performances of SimpleBaseline-a and mobileNetv1-SimpleBaseline.

Network	Parameter	Computational complexity (GFLOPs)	Inference time (ms)	PCK@40pixel(%)
SimpleBaseline-a	8.75 M	6.49	14.73	87.4
MobileNetv1-SimpleBaseline	9.50 M	14.07	18.17	90.9

TABLE 6 Comparison of model performance after distillation.

Network	Parameter	Computational complexity (GFLOPs)	Inference time (ms)	PCK@40pixel(%)
SimpleBaseline-lite	8.75 M	6.49	14.78	89.4
SimpleBaseline-a	8.75 M	6.49	14.73	87.4
MobileNetv1-SimpleBaseline	9.50 M	14.07	18.17	90.9

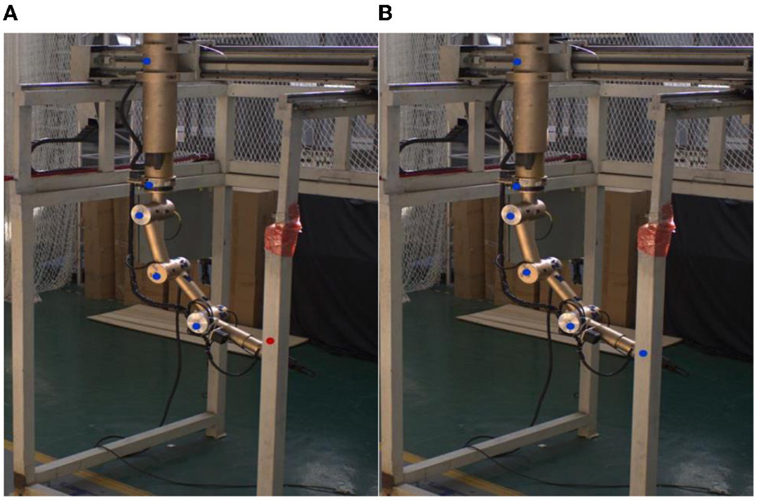


FIGURE 17
Visualization of model prediction results. **(A)** Before the "tutoring" by the teacher model. **(B)** After the "tutoring" by the teacher model.

the student model has an enhanced ability to extract difficult-to-extract feature information because the teacher model can give the student model extra supervision due to its excellent ability to extract global spatial information. Pose distillation has improved the ability of the student model to detect the key points of the teleoperation power manipulator.

5. Experiment

In this section, we selected 10 arbitrary pose images of the teleoperation power manipulator during its operation. Simultaneously, we recorded the readings from the demonstrator of the teleoperation power manipulator. These demonstrator readings serve as the true values for our measurements. Our measurement objectives encompass seven evaluation objects: the translational distance along the x-axis, the translational distance along the y-axis, the translational distance of the shoulder, the rotation angle of the upper arm, the rotation angle of the forearm, the rotation angle of the wrist, and the translational distance of the wrist. To assess the accuracy of our measurements, we utilized the errors associated with each evaluation object in every image as our evaluation indicators. In Experiment 1, the improved dilated-FCOS and SimpleBaseline-lite were used for pose estimation of the teleoperation power manipulator. In Experiment 2, the FCOS and SimpleBaseline were used to initialize the network with training weights through the same optimization method.

TABLE 7 Teleoperation power manipulator pose estimation experiment.

Error	Model	
	Dilated-FCOS + SimpleBaseline-lite	FCOS + SimpleBaseline
Translational distance along the x-axis/cm	6.27	6.36
Translational distance along the y-axis/cm	6.31	6.25
Translational distance of the shoulder/cm	4.32	4.34
Rotation angle of the upper arm/°	0.63	0.67
Rotation angle of the forearm/°	0.53	0.52
Rotation angle of the wrist/°	0.56	0.52
Translational distance of the wrist/cm	4.31	4.35

The pose estimation performances of different algorithms are shown in Table 7. The improved dilated-FCOS + SimpleBaseline-lite algorithm is superior to the FCOS + SimpleBaseline algorithm in some tasks, such as translation along the x-axis, translation of the shoulder, the rotation angle of the upper arm, and translation of the wrist, because the improved dilated-FCOS achieves the stable detection of the position of the teleoperation power manipulator by introducing a dilated encoder based on the characteristics of the teleoperation power manipulator and thus lays a good foundation for the subsequent pose estimation task. Other tasks show no significant differences between the two algorithms, which indicates that model weight reduction and pose distillation of SimpleBaseline have not significantly affected the model precision. However, in terms of computational speed, the average frame rate of the improved dilated-FCOS + SimpleBaseline-lite algorithm reaches 5.8 fps, while that of the original FCOS + SimpleBaseline-lite algorithm reaches ∼4.3 fps, which is 74% of that of the former. The results show that the pose estimation algorithm proposed in this paper has better performance in the teleoperation power manipulator pose estimation task than the FCOS + SimpleBaseline algorithm.

6. Conclusion

In this paper, the camera-based methods for target detection and pose estimation of teleoperation power manipulator is studied. The dilated-FCOS algorithm is proposed based on the FCOS algorithm and the scale of the teleoperation power manipulator. The shallow feature maps (P3, P4) of FCOS are discarded here to improve the detection speed of the FCOS, and the P7 feature layer of FCOS is discarded to improve the real-time performance of the network model. Model pruning is used to improve the real-time performance of the dilated-FCOS teleoperation power manipulator target detection model. To improve the detection speed for the key points of the teleoperation power manipulator, MobileNetv1 was selected as the backbone network based on the study of the SimpleBaseline algorithm and the comparison between keypoint detection precision and model inference speed of different lightweight backbone networks. To further optimize the inference speed of the model, the upsampling module was subjected to channel compression and pose distillation.

Our future work is as follows:

(1) The paper employs a motion capture system that relies on hand-eye calibration and an extrinsic calibration method for industrial cameras to track the movement of a teleoperation power manipulator. However, it is important to note that the current motion capture system may not be easily applicable in general scenarios. As a suggestion for future research, it would be beneficial to explore calibration methods that provide better generality and higher accuracy, addressing the limitations of the current approach.

(2) Model training is a critical aspect of supervised deep learning, where the quantity of training samples plays a significant role. However, in regular practice, the amount of available data is often limited. To address this limitation, future work can explore the utilization of simulation data from various scenarios to enhance the generalization ability of the model.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

LX: conceptualization, methodology, software, investigation, formal analysis, and writing—original draft. JH: data curation and writing—original draft. YL: visualization, investigation, and review and editing. JG: conceptualization, funding acquisition, resources, supervision, and writing—review and editing. All authors contributed to the article and approved the submitted version.

Funding

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation (2021B1515120009 and 2022A1515012004) and Project of Department of Education of Guangdong in China (2021ZDJS083).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Andriluka, M., Pishchulin, L., Gehler, P., and Schiele, B. (2014). "2D human pose estimation: new benchmark and state of the art analysis," in *IEEE Conference on Computer Vision and Pattern Recognition* (Columbus, OH: IEEE).

Assem, D., Christophe, S. V., and Fiona, M. (2014). Hot cell robot. *Nuc. Engin. Int.* 59, 30–31.

Chen, Q., Wang, Y., Yang, T., Zhang, X., Cheng, J., Sun, J., et al. (2021). "You only look one level feature," in *IEEE International Conference on Computer Vision and Pattern Recognition* (Montreal, QC: IEEE), 13034–13043. doi: 10.1109./CVPR46437.2021.01284

Coppelia Robotics GmbH. (2022). *V-REP Simulator*. Available online at: http://www.coppeliarobotics.com

Ferreira, J., Vale, A., and Ribeiro, I. (2012). Localization of cask and plug remote handling system in ITER using multiple video cameras. *Fusion Engin. Design*. 88:1992–1996. doi: 10.1016/j.fusengdes.10008

Heindl, C., Zambal, S., Ponitz, T., Pichler, A., and Scharinger, J. (2019). "3D robot pose estimation from 2D images," in *IEEE International Conference on Service Operations and Logistics, and Informatics* (Zhengzhou: IEEE), 95–99.

Henderson, P., and Ferrari, V. (2017). "End-to-end training of object class detectors for mean average precision," in *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20–24, 2016, Revised Selected Papers, Part V 13*. (Springer International Publishing), pp. 198–213. doi: 10.1007./978-3-319-54193-8_13

Hinton, G., Vinyals, O., and Dean, J. (2014). Distilling the knowledge in a neural network. *International Conference on Neural Information Processing Systems*, 38–39.

Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., et al. (2020). "Searching for mobileNetV3," in *IEEE/CVF International Conference on Computer Vision* (Seattle, WA: IEEE). doi: 10.1109/ICCV.2019.00140

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., et al. (2017). "MobileNets: efficient convolutional neural networks for mobile vision applications," in *IEEE International Conference on Computer Vision and Pattern Recognition* (Venice: IEEE), 6812–6820.

Kehl, W., Manhardt, F., Tombari, F., Ilic, S., and Navab, N. (2017). 6D: making RGB-based 3D detection and 6D pose estimation great again. *IEEE International Conference on Computer Vision*, 1530-1538. doi: 10.1109./ICCV.2017.169

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Adv. Neural. Inform. Proc. Sys*. 25, 1097–1105. doi: 10.1145/3065386

Lionel, M., Delphine, K., and Pierre, G. (2018). Virtual reality: lessons learned from west design and perspectives for nuclear environment. *Fusion Engin. Design.* 136(PT.B), 1337–1341. doi: 10.1016/j.fusengdes.05004

Liu, J., Lu, K., Pan, H., Cheng, Y., Zhang, T., Yao, Z., et al. (2020). Vision-based tile recognition algorithms for robot grasping task in EAST. *Fusion Engin. Design*. 152, 111422. doi: 10.1016/j.fusengdes.2019.111422

Liu, X., Yu, C., and Yang, D. (2018). "Inverted residuals and linear bottlenecks: mobile networks for classification, detection and segmentation," in *IEEE International Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 4025–4035. doi: 10.1109/CVPR.2018.00474

Liu, Z., Li, J., Shen, Z., Huang, G., Yan, S., Zhang, C., et al. (2017). Learning efficient convolutional networks through network slimming. *IEEE International Conference on Computer Vision*, 2755–2763.

Ma, N., Zhang, X., Zheng, H. T., and Sun, J. (2018). "V2: practical guidelines for efficient CNN architecture design," in *European Conference on Computer Vision* (Munich), 122–138. doi: 10.1007/978-3-030-01264-9_8

Maruyama, T., Aburadani, A., Takeda, N. (2014). Robot vision system RandD for ITER blanket remote-handling system. *Fusion Engin. Design* 89, 2404–2408. doi: 10.1016/j.fusengdes.01004

Mišeikis, J., Brijacak, I., Yahyanejad, S., Glette, K., Elle, O. J., Torresen, J., et al. (2018a). Transfer learning for unseen robot detection and joint estimation on a multi-objective convolutional neural network. IEEE International Conference on Intelligence and Safety for Robotics, 337–342. doi: 10.1109/IISR.2018.8535937

Mišeikis, J., Brijacak, I., Yahyanejad, S., Glette, K., Elle, O. J., Torresen, J., et al. (2018b). Multi-objective convolutional neural networks for robot localisation and 3D position estimation in 2D camera images. *IEEE International Conference on Ubiquitous Robots*, 597–603. doi: 10.1109/URAI.2018.8441813

Motive Optical motion capture software. (2023). Available online at: https://www.optitrack.com/software

Ning, X., Duan, P., Li, W., and Zhang, S. (2020). Real-time 3D face alignment using an encoder-decoder network with an efficient deconvolution layer. *IEEE Sig. Process. Lett.* 27, 1944–1948. doi: 10.1109/LSP.2020.3032277

O'Donovan, K. J., Kamnik, R., O'Keeffe, D. T., and Lyons, G. M. (2006). An inertial and magnetic sensor based technique for joint angle measurement. *J. Biomech*. 40, 2604–2611. doi: 10.1016/j.jbiomech.12,010

Oosterhout, J. V., Abbink, D. A., Koning, J. F. (2012). Haptic shared control improves hot cell remote handling despite controller inaccuracies. *Fusion Engin. Design* 88, 2119–2122. doi: 10.1016/j.fusengdes.11006

Pezhman, L., and Saeed, S. A. (2011). novel approach to develop the control of Telbot using ANFIS for nuclear hotcells. *Annals Nuc. Energy* 38, 2156–2162. doi: 10.1016/j.anucene.06021

Qiu, Q., Gu, K., Wang, P. (2016). Hand-eye coordinative remote maintenance in a tokamak vessel. *Fusion Engin. Design* 104, 93–100. doi: 10.1016/j.fusengdes.01006

Redmon, J., and Farhadi, A. (2017). YOLO9000: better, faster, stronger. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; Piscataway: IEEE, 7263–7271. doi: 10.1109/CVPR.2017.690

Ren, S., He, K., Girshick, R., and Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. NIPS 3, 7031. doi: 10.1109./TPAMI.2016.2577031

Ribeiro, L. G., Suominen, O. J., Peltonen, S., Morales, E. R., and Gotchev, A. (2020). Robust vision using retro reflective markers for remote handling in ITER. *Fusion Engin. Design* 161, 112080. doi: 10.1016/j.fusengdes.2020.112080

Simon, M., Rodner, E., and Denzler, J. (2016). "ImageNet pre-trained models with batch normalization," in *IEEE International Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 3124–3135.

Sun, K., Xiao, B., Liu, D., and Wang, J. (2017). Deep high-resolution representation learning for human pose estimation. *IEEE International Conference on Computer Vision and Pattern Recognition*, 5686–5696.

Szegedy, C., Ioffe, S., and Vanhoucke, V. (2016). "Inception-v4, inceptionresnet and the impact of residual connections on learning," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (San Francisco, CA: AAAI Press). doi: 10.1609./aaai.v31i1.11231

Tian, Z., Shen, C., Chen, H., and He, T. F. C. O. S. (2020). Fully convolutional one-stage object detection. *IEEE International Conference on Computer Vision*, 9626–9635. doi: 10.1109./ICCV.2019.00972

Toshev, A, and Szegedy, C. (2014). DeepPose: human pose estimation via deep neural networks. *IEEE Conference on Computer Vision and Pattern Recognition*, 1653–1660. doi: 10.1109./CVPR.2014.214

Wei, S. E., Ramakrishna, V., Kanade, T., and Sheikh, Y. (2016). Convolutional Pose Machines. *IEEE Conference on Computer Vision and Pattern Recognition*, 4724-4732. doi: 10.1109/CVPR.2016.511

Wu, G., He, F., Zhou, Y., Jing, Y., Ning, X., Wang, C., et al. (2022). Age-compensated makeup transfer based on homologous continuity generative adversarial network model. *IET Comp. Vis*. 4, 2138. doi: 10.1049/cvi2.12138

Xiao, B., Wu, H., and Wei, Y. (2018). "Simple baselines for human pose estimation and tracking," in *European Conference on Computer Vision (ECCV)* (Munich). doi: 10.1007./978-3-030-01231-1_29

Zeng, L., Yuan, F., Song, D., and Zhang, R. (1999). A two-beam laser triangulation for measuring the position of a moving object. *Optics Lasers Engin*. 31, 445–453. doi: 10.1016/S0143-8166(99)00043-3

Zhang, S., Wang, Z., Li, C., Zhao, Y., Feng, C., Dai, M., et al. (2022). Study on neutron shielding performance of hot cell shielding door for nuclear power plant. *Ann. Nuc. Energy* 166, 108752. doi: 10.1016/j.anucene.2021.108752

Zhang, T. Y., Li, S., and Zhang, W. J. (2021). Design of power manipulator for hot cell facility. *2021 IEEE International Conference on Robotics and Biomimetic*, 458–462. doi: 10.1109./ROBIO54168.2021.9739245

Zheng, G., Minzhong, Q., Yong, C. H., and Yuntao, S. (2015). Conceptual layout design of CFETR Hot cell facility. *Fusion Engin. Design* 100, 280–286. doi: 10.1016/j.fusengdes.06088

TYPE Original Research
PUBLISHED 08 August 2023
DOI 10.3389/fnbot.2023.1205370

Check for updates

OPEN ACCESS

EDITED BY
Long Jin,
Lanzhou University, China

REVIEWED BY
Stefano Marrone,
University of Naples Federico II, Italy
Shuqiang Wang,
Chinese Academy of Sciences (CAS), China

*CORRESPONDENCE
Xingyuan Chen
✉ chxy302@vip.sina.com

RECEIVED 13 April 2023
ACCEPTED 13 July 2023
PUBLISHED 08 August 2023

CITATION
Qin R, Wang L, Du X, Xie P, Chen X and Yan B (2023) Adversarial robustness in deep neural networks based on variable attributes of the stochastic ensemble model.
Front. Neurorobot. 17:1205370.
doi: 10.3389/fnbot.2023.1205370

Adversarial robustness in deep neural networks based on variable attributes of the stochastic ensemble model

Ruoxi Qin[1], Linyuan Wang[2], Xuehui Du[1], Pengfei Xie[1], Xingyuan Chen[2]* and Bin Yan[1]

[1]Henan Key Laboratory of Imaging and Intelligent Processing, PLA Strategy Support Force Information Engineering University, Zhengzhou, Henan, China, [2]PLA Strategy Support Force Information Engineering University, Zhengzhou, Henan, China

Deep neural networks (DNNs) have been shown to be susceptible to critical vulnerabilities when attacked by adversarial samples. This has prompted the development of attack and defense strategies similar to those used in cyberspace security. The dependence of such strategies on attack and defense mechanisms makes the associated algorithms on both sides appear as closely processes, with the defense method being particularly passive in these processes. Inspired by the dynamic defense approach proposed in cyberspace to address endless arm races, this article defines ensemble quantity, network structure, and smoothing parameters as variable ensemble attributes and proposes a stochastic ensemble strategy based on heterogeneous and redundant sub-models. The proposed method introduces the diversity and randomness characteristic of deep neural networks to alter the fixed correspondence gradient between input and output. The unpredictability and diversity of the gradients make it more difficult for attackers to directly implement white-box attacks, helping to address the extreme transferability and vulnerability of ensemble models under white-box attacks. Experimental comparison of *ASR-vs.-distortion curves* with different attack scenarios under CIFAR10 preliminarily demonstrates the effectiveness of the proposed method that even the highest-capacity attacker cannot easily outperform the attack success rate associated with the ensemble smoothed model, especially for untargeted attacks.

KEYWORDS

deep neural network, adversarial robustness, stochastic ensemble, random smoothing, cyberspace security

1. Introduction

Deep learning techniques have been successfully applied in various computer vision applications, ranging from object detection (Ren et al., 2016) and image classification (Perez and Wang, 2017) to facial recognition (Parkhi et al., 2015) and autonomous driving (Bojarski et al., 2014) and even in medical computer-aided diagnosis (Hu et al., 2020; You et al., 2022). In these application scenarios, deep learning can be used as an enhancement technique for real data as an artificial intelligence generated content (AIGC) technique to improve performance on the one hand, and as a tool to generate false data to degrade the performance of the model on the other. However, with the increasing use of deep neural networks (DNNs) in various application areas, such as facial recognition technology, for encryption applications, autonomous driving technology for road safety, and computer-aided diagnosis for life safety, there is an urgent need to principally ensure effective defense against security threats, not just the good performance.

The studies on adversarial samples reveal the extreme vulnerability of deep networks, making the study of their robustness even more urgent for security applications. In, Szegedy et al. (2014) discovered that the input-to-output mappings learned by DNNs are generally discontinuous so that even small perturbations in some network inputs can lead to high misclassification errors, which are known as adversarial samples. As a result, many adversarial learning methods similar to cyberspace security games have been developed for both the attack and defense sides. Research on attack and defense in DNN primarily focuses on adversarial samples because of their proactive role in attack and defense games (Akhtar and Mian, 2018; He et al., 2020).

The development of attack methods is constantly intertwined with the proposal of defense methods. Both types of methods act as opposing sides in a competitive game, developed in a mutually promoting and closely reciprocal process. Certified defense methods are supported by rigorous theoretical security guarantees that obtain a robustness radius under the Lp distortion constraint (Fischetti and Jo, 2017). Nevertheless, these certified defense methods are still not widely used in DNN architectures on big data through exact or conservative approaches. More flexible and effective defense methods are empirical methods based on assumptions and experimental results (Papernot et al., 2016; Lakshminarayanan et al., 2017; Kurakin et al., 2018). Although empirical defense methods are convenient, they have practical limitations in their applicability, which may result in attackers generating more challenging adversarial samples to break the defense.

The rapid development of attack algorithms and extensive research on empirical defenses eventually led to the game of attack and defensive in deep learning files. For example, the distillation method (Papernot et al., 2016) which uses gradient shielding to prevent white-box attacks, is not effective against the CW attack (Carlini and Wagner, 2017). The model ensemble method (Lakshminarayanan et al., 2017) was initially proposed as a defense method but has been found to be ineffective (He et al., 2017) and is now commonly used as an attack method to improve the transferability of adversarial samples (Tramèr et al., 2018). The nature and wide applicability of empirical defense methods have sparked intense competition with attack methods. However, defense methods are primarily passive.

According to theoretical developments in cybersecurity, the two sides in a competitive game without a strongly secure defense method will eventually reach a Nash equilibrium (Attiah et al., 2018). To address this challenge, generalized robust-control defense methods, such as moving target defense (MTD) (Jajodia et al., 2011) and dynamic defense model (DDM) (Wu et al., 2019; Wu, 2020), have been proposed with probabilistic formulations of the network attributes. The inherent randomness and unpredictability of the system make it more difficult for the attacker to detect, highlighting the importance of the same defense approach applied in DNNs. Recent research on adversarial robustness indicates that adversarial examples are inevitable for DNNs. This article starts from the premise of learning from the development experience of cybersecurity under the current technical levels and treating the classification problem based on deep neural networks as a whole system rather than a single model. In the case where effective adversarial samples mostly depend on specific model information while the adversarial transferability needs to be improved, this article proposes an attribute-based stochastic ensemble model using the DDM ideology to combine randomness with model diversity. In the proposed method, the ensemble quantity, network architecture, and smoothing parameters are used as ensemble attributes to dynamically change it before each inference prediction request. As shown in Figure 1, these variable attributes of the ensemble model represent a more active and generalized defense approach to overcome the limitations of empirical and deterministic defense at the current stage. In summary, the main contributions of our study are as follows:

(1) Facing the endless arms race of adversarial attack and defense, this article proposes an attribute-based stochastic ensemble model using the DDM ideology to combine randomness with model diversity. A more diverse collection of heterogeneous and redundant models is created for the ensemble, accounting for variations in ensemble attributes and dynamically changing structures for each inference prediction request at the model level, hoping to further change the passive position of the defense at this stage.
(2) For the robustness evaluation of the proposed method, this article considers the attack and defense game idea as a starting point, assuming that the attacker knows the defense strategy, and simulates a series of possible adversarial game processes for a more comprehensive evaluation. The different capabilities of the attack scenario are set up and the potential defense risks are assessed using attack success rate versus distortion (*ASR-vs.-distortion curves*) based on Monte Carlo simulations.
(3) We analyze different robustness results under attack scenarios and algorithms with various capabilities and identify important conditions for the proposed method to exert its advantages in practice. The experimental results under CIFAR10 show that even the most capable attacker is unable to outperform the best result under current random-based methods, demonstrating the effectiveness of the proposed method in attack and defense games.

2. Related work

2.1. Defense method based on input randomization

Recently, theoretical guarantees for the robustness of DNNs have been gradually combined with relevant aspects of cybersecurity. Random smoothing was originally proposed based on the intention of differential privacy (Lecuyer et al., 2019) from cyberspace defense methods to prevent the attackers from obtaining exact gradient information by adding random noise to the input image during training and testing (Cohen et al., 2019; Lecuyer et al., 2019; Li B. et al., 2019). Random self-ensemble (RSE) (Liu et al., 2018) and Smoothed WEighted ENsemble (SWEEN) (Liu et al., 2020) improve the adversarial robustness by combining the randomness properties in the case of the ensemble.

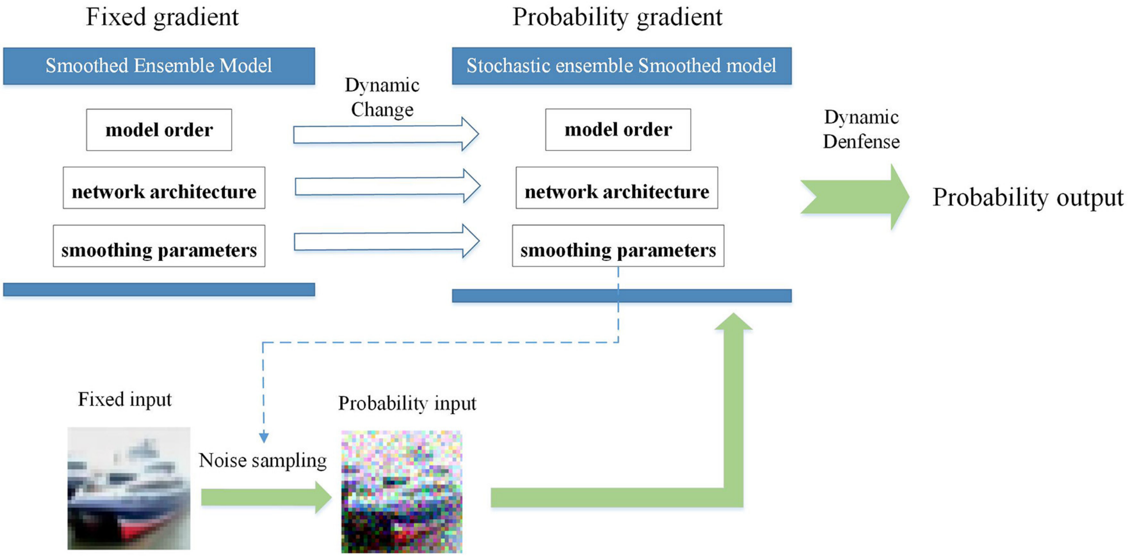


FIGURE 1
A graphical illustration of the proposed variable attributes of ensemble strategy for adversarial robustness of deep neural networks.

Unlike these previous studies, this study is inspired by the DDM ideology in cyberspace security and sets the model parameters on which the attack conditions directly depend as the objects of randomization to further improve the adversarial robustness under ensemble conditions.

2.2. Defense methods based on diversified ensemble networks

In addition to the gradient shielding effect of the random smoothing, the robustness provided by ensemble models also depends on the diversity of the sub-model (Lakshminarayanan et al., 2017). Constraints on the gradient diversity of sub-models mostly depend on empirical conclusions about the diversity of model architecture (Kurakin et al., 2018) or the training hyperparameters (Wenzel et al., 2020) and gradient diversity between sub-models (Pang et al., 2019). Unlike the fixed ensemble of diverse sub-models in these methods, this study uses the empirical conclusions of model attributes to contrast the diverse sub-models. By randomly selecting these attributes, this method combines diversity and randomization characteristics to improve adversarial robustness under the ensemble condition.

2.3. Adversarial samples and robustness evaluation

Attack algorithms can be divided into white-box and black-box methods based on their capabilities (Akhtar and Mian, 2018). White-box methods rely on full knowledge of the network gradients. The fast gradient sign method (FGSM) (Goodfellow et al., 2015) is a basic and effective method that generates adversarial samples by adding the sign reverse of the gradient to the original images. Based on attack performance and transferability, iteration-based approaches include the basic iterative method (BIM) (Kurakin et al., 2016), momentum iterative method (MIM) (Dong et al., 2018), and projected gradient descent method (PGD) (Madry et al., 2018). In contrast, black-box attackers have no knowledge of the network gradients that can be divided into query-based and transfer-based methods. The query-based method achieves gradient estimation by querying the output of the target model including natural evolution strategies (NES) (Ilyas et al., 2018), simultaneous perturbation stochastic approximation (SPSA) (Uesato et al., 2018), and NATTACK (Li Y. et al., 2019). The transfer-based method generates adversarial samples by constructing substitution models, usually using the ensemble model constructed by normally trained sub-models (Tramèr et al., 2018) or shadow model (Zhang et al., 2022). In previous studies, different adversarial sample generation algorithms can verify the different performances of the defense method from different perspectives. Unlike the previous single analysis of the defense capability under optimal attack algorithms, this study considers the game-like nature of the attackers and designs more diverse attack and defense scenarios under random conditions to fully verify the effectiveness of the proposed method.

3. Materials and methods

This study focuses on the image classification task of CIFAR10 (Krizhevsky and Hinton, 2009) for preliminary verification. Section 3.1 first introduces the basic method of random smoothing and shows the relationship with the proposed stochastic ensemble model to theoretically demonstrate that the proposed method achieves a certified robust radius no less than the state-of-the-art (Liu et al., 2020) under the random conditions. Furthermore, the empirical diversity requirement between sub-models in the ensemble is characterized by attribute-based heterogeneous redundant models to improve the robustness of the stochastic ensemble model in Section 3.2. Finally, Section 3.3 outlines the strategy for a stochastic ensemble approach with variable attributes.

3.1. Preliminaries of stochastic ensemble modeling

Let the random smoothing model g be trained by a basic classifier f by sampling, adding the noise $\delta \sim N\left(0, \sigma^2 I\right)$ to the input images and minimizing the corresponding classification losses (Cohen et al., 2019; Lecuyer et al., 2019; Li B. et al., 2019). For the model prediction in the training and testing process, the output of random smoothing model g is defined as a mathematical equation as follows:

$$g\left(x\right) = E_{\delta \sim N\left(0,\sigma^2 I\right)}\left[f\left(x+\delta\right)\right] \tag{1}$$

An ensemble model f_{ens} containing K models obtains the final prediction by summing the function outputs of the individual candidate models. The mathematical representation of the ensemble model can be written as follows:

$$f_{ens}\left(x, \theta\right) = \sum_{k=1}^{K} f\left(x, \theta_k\right) \tag{2}$$

The SWEEN approach creates an ensemble-smoothed model with a weight parameter ω for each model, which improves the provable robustness radius (Liu et al., 2020). In terms of the probability distribution of the input noise, the predicted output of the SWEEN model is given by a mathematical expectation operator as follows:

$$\begin{aligned} SWEEN = E_{\delta}\left[\sum_{k=1}^{K} \omega_k f\left(x+\delta;\theta_k\right)\right] &= \sum_{k=1}^{K} \omega_k E_{\delta}\left[f\left(x+\delta;\theta_k\right)\right] \\ &= \sum_{k=1}^{K} \omega_k g\left(x;\theta_k\right) \end{aligned} \tag{3}$$

The constant weight parameters ω of the candidate models are independent of the SWEEN model output and can be optimized as ω^*. Unlike SWEEN, the ensemble attributes of the proposed stochastic ensemble model (SEM) are randomly adjusted to dynamically structure the ensemble model at each time inference prediction request making the output of candidate models in SEM have an additional mathematical expectation in terms of probability of occurrence. However, the probability of occurrence of a particular candidate model under the SEM is assumed to be determined by the expectation $E\left(f_k\right)_{occurrence} = \omega_k$ and statistically independent of the prediction expectation. Therefore, as shown in Equation (4), the stochastic ensemble and SWEEN models can be equivalent in terms of output expectations. The theoretical improvement of the robustness radius by the SWEEN model (Liu et al., 2020) is a special case of the SEM. By controlling the probability of the occurrence of sub-models, the SEM can theoretically achieve well-certified robustness. However, more importantly, such changes based on the model level improve the dynamic properties of the ensemble and achieve a more generalized dynamic change of the model gradient in each inference prediction:

$$\begin{aligned} &\mathrm{SEM} = E\left[\sum_{k=1}^{K} f_k\left(x+\delta;\theta_k\right)\right] = \sum_{k=1}^{K} E\left(f_k\right)_{apparence} \\ &\quad\times E\left[f_k\left(x+\delta;\theta_k\right)\right] \\ &\mathrm{SEM}{=} \sum_{k=1}^{K} \omega_k^* E\left[f_k\left(x+\delta;\theta_k\right)\right]{=} \sum_{k=1}^{K} \omega_k E\left[f_k\left(x+\delta;\theta_k\right)\right] \\ &= \mathrm{SWEEN\ when\ } \omega_k = \omega_k^* \end{aligned} \tag{4}$$

3.2. Attributes-based heterogeneous redundant models

The application of random input to the sub-model parameters in SWEEN (Liu et al., 2020) improves the certified robustness of the ensemble. The analysis in Section 3.1 has shown that these sub-models can also serve as a random condition, expanding randomness at the model level without compromising the certified robustness. According to previous empirical defense conclusions, the diversity between sub-models enhances the robustness of the ensemble condition (Pang et al., 2019; Wenzel et al., 2020). Moreover, diversity is also the DDM property in cybersecurity (Wu et al., 2019). Therefore, the first step for the proposed variable attribute-based SEM is a collection of heterogeneous redundant sub-models. In addition to the diversity of the model architectures (Kurakin et al., 2018), different hyperparameters for optimizing the sub-models can also have different effects on the convergence of the gradient (Wenzel et al., 2020). Random smoothing hyperparameters for a variety of noise parameters in training further enhance model redundancy and diversity within the same architecture. The proposed SEM uses network architecture, depth, and width as well as smoothing parameters as variable ensemble attributes. In Section 4.5, we present detailed experimental results on the influence of model architecture and other parameters.

The heterogeneous redundant model collection is obtained by separately training a smoothed model on the CIFAR10 dataset (Krizhevsky and Hinton, 2009; Hendrycks et al., 2019). The variable ensemble attributes in this study include architectures of different depths and widths. Table 1 shows the *approximated certified accuracy* (ACA) of the predictive performance of each sub-model. The models marked in red did not meet performance requirements and were excluded from subsequent experiments. Although some simple models, such as AlexNet and shallow VGG, were unable to achieve stable smoothed prediction, unsmoothed models were used for the SEM. The experimental results in Section 4.5 further demonstrate that the heterogeneity of the model collection plays a crucial role in the robustness of the stochastic ensemble.

3.3. Stochastic ensemble with variable attributes

In a model ensemble, temporal gradient variations result from attribute-based gradient changes in each smoothed model. This article proposes a stochastic ensemble strategy based on

TABLE 1 Heterogeneous redundant model collection on CIFAR10.

Model architecture	Smoothing parameter σ			Model architecture	Smoothing parameter σ		
	0.25	0.75	1.5		0.25	0.75	1.5
DenseNet (Gao et al., 2017)				**VGG** (Simonyan and Zisserman, 2014)			
DenseNet100 (95.5)	94.03	89.96	83.56	VGG11 (92.1)	9.99	80.11	20.88
DenseNet121 (94.1)	91.23	87.01	82.08	VGG13 (94.3)	65.67	10.0	61.18
DenseNet161 (94.2)	92.31	87.88	82.80	VGG16 (93.9)	9.99	9.99	9.99
DenseNet169 (94.0)	91.29	87.96	81.11	VGG19 (93.3)	91.83	87.50	81.74
WRN (Zagoruyko and Komodakis, 2016a) (96.2)	91.78	90.23	83.43	**AlexNet** (Krizhevsky et al., 2017) (77.2)	9.99	9.99	9.99
ResNet (He et al., 2016)				**InceptionV3** (Szegedy et al., 2016) (93.8)	91.91	86.86	80.38
ResNet18 (93.3)	90.49	86.63	80.15	**MobileNetV2** (Sandler et al., 2018) (94.2)	88.91	84.74	77.35
ResNet34 (92.9)	91.20	87.20	81.76	**ResNext** (Xie et al., 2017) (96.2)	93.12	88.70	80.62
ResNet50 (93.9)	91.16	86.29	80.28	**GoogleNet** (Szegedy et al., 2015) (92.7)	91.63	87.61	80.64

heterogeneous redundant models, where each prediction is made by the stochastic selection of ensemble attributes. The randomness of the model attributes reflects SEM randomness, which varies in the frequency of the ensemble quantity, network architecture, and smoothing parameters when multiple requests for gradient or output information are made. The model randomly selects the number of sub-models for the ensemble. Once the number of ensemble models has been determined, the model stochastically selects the model architecture from Table 1. Next, it randomly selects various parameters of the selected model architecture, such as network depth and smoothing parameters. Finally, the ensemble model is determined based on these stochastic ensemble attributes. Algorithm 1 provides a detailed explanation of the selection process for this method.

```
Require: Image x for classification, K-ensemble quantity, f-model architecture, $\delta$-smoothing parameter; $f_k(x+\delta)$-model source output before softmax
Ensure: $output_{ensemble}$-softmax operation of ensemble model
While inference prediction request for one user do
    Randomly determine the model quantity K for the ensemble;
    Randomly select the number of model architectures $f$ according to model quantity K;
    Randomly select different smoothing parameters $\delta$ for each model architecture, the sub-model of ensemble is determined by $f_k$ finally;
    $source_{ensemble} \leftarrow 0$
    for each $k \in [1, K]$ do
        $source_{model} \leftarrow f_k(x+\delta)$
        $source_{ensemble} \leftarrow source_{ensemble} + source_{model}$
    end for
    $output_{ensemble} \leftarrow softmax(source_{ensemble})$
end while
```

Algorithm 1. Framework of the stochastic ensemble for the defense system.

Figure 2 shows a flowchart of the stochastic ensemble strategy. By incorporating the model architecture into ensemble attributes, each iteration of the ensemble incorporates gradient differences based on changes in the network architecture. In addition, network depth and smoothing parameters were used as ensemble attributes to increase ensemble diversity. The number of sub-models in each ensemble iteration is relatively small [set as (1–4) in this article] compared to all of the model collections to ensure gradient differentiation. On the one hand, a larger number of sub-models sets in each ensemble iteration will reduce the ensemble diversity and gradient variations. On the other hand, a large number of sub-models sets in the ensemble will lead to improved transferability of adversarial samples generated from a possible white-box attack for a single ensemble iteration. For probabilistic ensembles, allowing a single model in the stochastic state does not affect the mathematical expectation of the prediction, but ensures a diversity gradient change in each ensemble iteration. The attribute of the ensemble quantity plays a key role and has an important impact on robustness, which will be discussed in detail in Section 4.5.

The SEM introduces the dynamic nature of DNNs through the stochastic selection of the ensemble attributes. The dynamic changes reflect the random distribution of input noise and probabilistic gradient information during each ensemble iteration. Essentially, the randomness of ensemble attributes shields the gradient information and increases the confusion under white-box and query-based black-box attacks.

4. Experiments and results

Currently, most single static models rarely consider both white-box and black-box attack robustness evaluation comprehensively but consider white-box attack robustness as the evaluation metric. The probabilistic gradient of the proposed SEM makes it difficult for attackers to fully discover the model parameter of each particular ensemble iteration. From the attackers' point of view, the more effective attack is no longer the white-box attack defined in the original evaluation but is based on the attacker's knowledge of the model collection to achieve the black-box attack or approximate white-box attack. This section comprehensively designs different knowledge of attacker against the SEM and comprehensively illustrate the potential and drawbacks of the proposed method. To define and evaluate the robustness under random conditions, the attack success rate is further defined as a potential risk by *ASR-vs.-distortion* curves (Dong et al., 2019) based on Monte Carlo simulations. For the conclusion of robustness, this section generally verified and evaluated adversarial robustness same as the definition in cyberspace security: the most capable attacker for SEM cannot easily outperform the best result under current random-based methods.

4.1. Attack success evaluation metrics based on empirical risk

The *ASR-vs.-distortion* curves are generated by an optimal search of the adversarial perturbation budget (Dong et al., 2019). Due to the random condition, the Monte Carlo simulation is used for approximate evaluation as in random smoothing (Cohen et al., 2019). Each adversarial sample x_{adv} is hard-predicted N times by the SEM, and the most predicted category is considered the output category with the highest probability. The baseline accuracy of the clean sample through this simulation is 93.4%. Compared with the according accuracy result of the single smoothing model in Table 1, there is no damage but even improvement for clean-sample prediction. The attack success rate with the adversarial sample x is given as follows:

$$Succ\left(C, A_{\varepsilon,p}\right) = \begin{cases} \frac{1}{N}\left(\sum_{n=1}^{N}\left(\sum_{k=1}^{K} g_k\left(A_{\varepsilon,p}\left(x\right)\right)\right)_{one\_hot}\right)_{\max} \neq y \text{ untargeted} \\ \frac{1}{N}\left(\sum_{n=1}^{N}\left(\sum_{k=1}^{K} g_k\left(A_{\varepsilon,p}\left(x\right)\right)_{one\_hot}\right)\right)_{\max} = y_t \text{ targeted} \end{cases} \tag{5}$$

The attack success probability is redefined as the proportion of Monte Carlo simulations in which each k-th iteration model g_k outputs the target category for the given adversarial sample $A_{\varepsilon,p}$ with a perturbation budget ε under the l_p norm. This probability is estimated using class count statistics obtained by one-hot encoding of the category probability vector, and then converting each predicted value to its equivalent probability using a probability conversion function. Such probabilities can be used in a two-sided hypothesis test that the attack success rate conforms to the binomial distribution $n_{succ} \sim Binomial\left(n_{succ} + n_{nonsucc}, \rho\right)$ as follows:

$$Succ\left(C, A_{\varepsilon,p}\right) = \begin{cases} \frac{1}{N}\left(\sum_{n=1}^{N}\left(\sum_{k=1}^{K} g_k\left(A_{\varepsilon,p}\left(x\right)\right)\right)_{one\_hot}\right)_{\max} \neq y \text{ or} \\ \frac{1}{N}\left(\sum_{n=1}^{N}\left(\sum_{k=1}^{K} g_k\left(A_{\varepsilon,p}\left(x\right)\right)\right)_{one\_hot}\right)_{\substack{\max \\ c\neq y}} \geq \alpha \text{ untargeted} \\ \frac{1}{N}\left(\sum_{n=1}^{N}\left(\sum_{k=1}^{K} g_k\left(A_{\varepsilon,p}\left(x\right)\right)\right)_{one\_hot}\right)_{\max} = y_t \text{ or} \\ \frac{1}{N}\left(\sum_{n=1}^{N}\left(\sum_{k=1}^{K} g_k\left(A_{\varepsilon,p}\left(x\right)\right)\right)_{one\_hot}\right)_{t} \geq \alpha \text{ targeted} \end{cases} \tag{6}$$

The abstention threshold α is a parameter used to limit the probability of returning an incorrect prediction in order to control potential empirical model risk (Hung and Fithian, 2016). A value of α directly affects the *ASR-vs.-distortion* curves. In this case, the threshold α is set at 0.3 to evaluate the random smoothing model.

4.2. Attack scenarios

In this section, the attacker's knowledge of the SEM attributes is discussed in detail and the attack scenarios are designed to fully characterize the robustness of the proposed method. By comparing the robustness evaluation results of attackers with different capabilities under the proposed method with the results of the contrast models, the attack scenarios are designed to discuss two aspects of robustness: first, under which attack capabilities is the proposed method most vulnerable and which is the most robust. This will help defenders to understand which attributes are important for protection. Second, whether the proposed method is robust enough such that even an attacker with the highest attack capability cannot easily exceed the attack success rate associated with the best contrast method (Athalye et al., 2018).

In the random condition, different attackers can have different degrees of knowledge about the model collection, but no knowledge about the current ensemble state. From an attack point of view, the attacker should use a white-box attack under expectation, a transfer-based attack under the substitution model, or a query-based black-box attack. The attacker's capabilities are determined by the knowledge of the model collection and the ensemble attributes, as outlined from high to low in Table 2. In the white-box attack under expectation, attackers A and B have full knowledge of model collection and are implemented as Expectation Over Transformation (EOT) attack method (He et al., 2017; Croce et al., 2022) white-box attack according to the different expectation estimation iteration. In the transfer-based attack under the substitution model, attackers C and D have partial knowledge of the model collection and are defined according to the different transfer strategies. In addition, attacker E uses the query-based black-box attack algorithm. The analysis of our experimental setup highlights the varying ability of the A–D attackers to approximate the gradient distribution expectation, which comprehensively illustrates the robustness of our method under more complicated conditions.

4.3. Experimental settings of competitive baseline methods

To verify the improvement of robustness, several ensemble methods were selected as baselines for comparison, including RSE (Liu et al., 2018), random smoothing (Liu et al., 2020), and the adaptive diversity promoting (ADP) (Pang et al., 2019). For the details of the experiment, both the random smoothing ensemble and baseline ensemble method used three different model architectures, namely, DenseNet100, ResNet50, and WRN, as shown in Table 1, which perform better on clean datasets. The parameters of the smoothed models were chosen as Gaussian noise with δ 0.25. Figure 3 shows that neither the ADP nor the

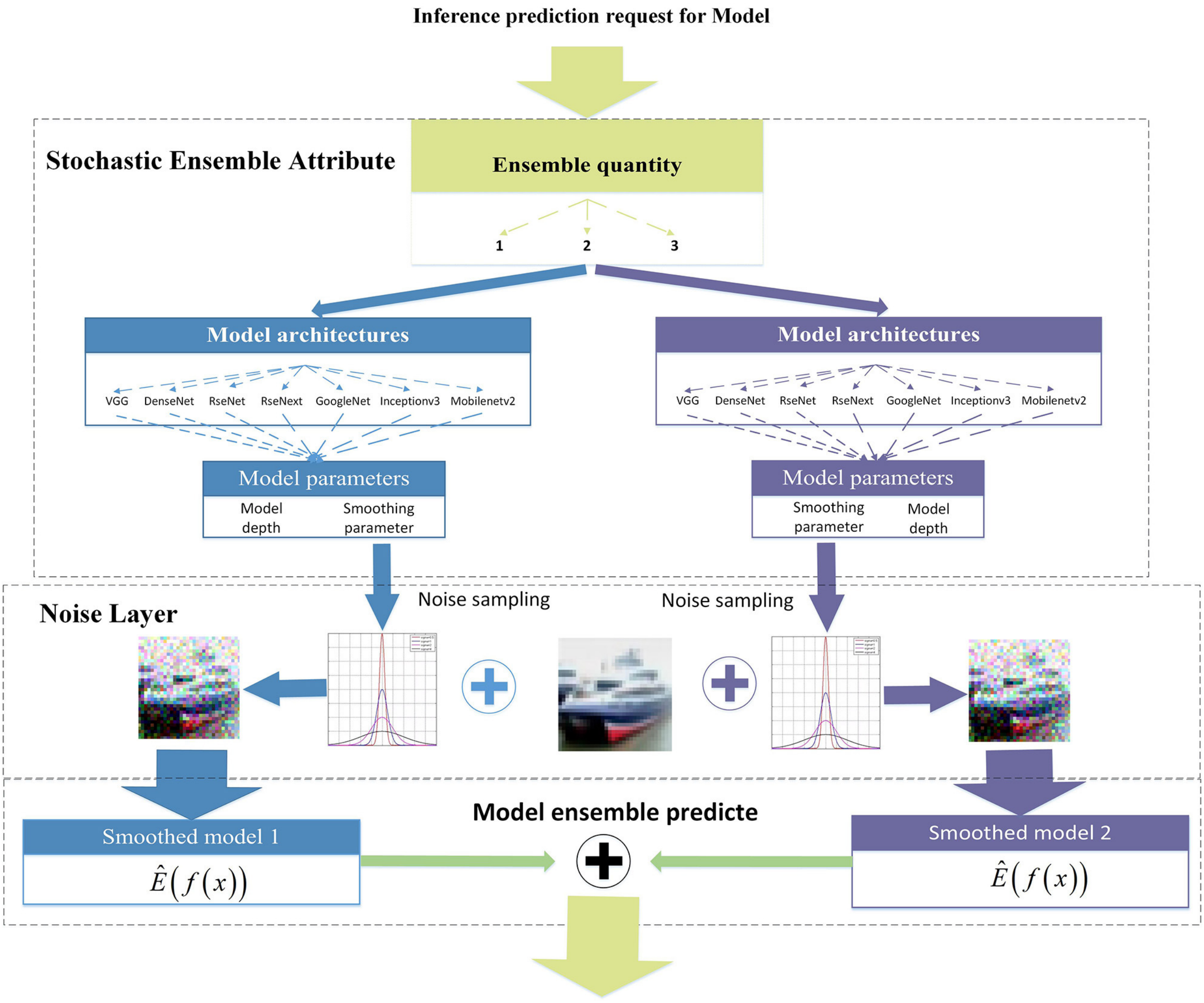


FIGURE 2
A flowchart of the stochastic ensemble smoothing strategy.

RSE methods outperform the ensemble-smoothed method. Among the defenses based on randomness and ensemble diversity, the ensemble smoothed model has SOTA results at this stage and structure as the contrast method F in attack scenarios. In a follow-up experiment, the random smoothing-related method with the best robustness is used as a contrast method (corresponding to the four curves of F, G, J, and K in the contrast methods as shown in Table 2) to demonstrate the performance of the proposed method for brief.

4.4. Robustness analysis based on the attack scenario

A comprehensive evaluation of adversarial robustness can be achieved by considering different combinations of attack capabilities, methods, targets, and perturbation constraints. Further attacks are carried out by the algorithm using three standard methods (BIM, MIM, and PGD) with attackers *A, B, C,* and *D* and contrast methods *F, G, H,* and *I*, respectively. In addition, NES and SPSA attacks were used in conjunction with contrast methods E, G, K, L, and M. For all *ASR-vs.-distortion* curves, the search step was set to 10 while the binary search step was set to 20. For the white-box attacks, the number of attack iterations of both the BIM and MIM was set to 20, while for the query-based black-box attacks, the maximum number of queries was set to 5000. The following experiments aim to evaluate the proposed methods and analyze the defense characteristics of dynamics under different attack scenarios set in Section 4.2.

4.4.1. Transfer-based and white-box attack analysis

Figure 4 shows the *ASR-vs.-distortion* curves for untargeted transfer-based attacks. *A, B, C,* and *D* represent different attack scenarios, while the contrast methods *F, G, H,* and *I* are shown

TABLE 2 The definition of the attacker's ability from high to low and the contrast method.

Attacker tag		Definition	Contrast method		Definition
White-box attack as EOT	Attacker A	The attacker has full knowledge of the model collection and can obtain ensemble attributes in real-time. However, they lack the ability to predict these attributes for the next ensemble iteration, where their best strategy is to implement the EOT attack on each ensemble iteration for the expectation of gradient.	Under White-box attack	Contrast method F	The ensemble smoothed model under a white-box attack
				Contrast method G	The single-smoothed model under a white-box attack
	Attacker B	The attacker has full knowledge of the model collection but cannot obtain ensemble attributes in real-time, where their one of the attack strategies is to implement an EOT attack on periodic ensemble iteration.		Contrast method H	The ensemble model under a white-box attack
				Contrast method I	The single model under a white-box attack
Transfer-based black-box attack	Attacker C	The attacker has knowledge of half of the models in the collection for the experiment. Their best attack strategy is to structure the alternative SEM model on known models as an EOT method for generalized adversarial samples.	Under Black-box attack	Contrast method J	The ensemble smoothed model under the black-box attack
				Contrast method K	The smoothed model under the black-box attack
	Attacker D	The attacker has knowledge of half of the models in the collection. Their more direct attack strategy is to use all the known models as an ensemble model to generate transfer adversarial samples.		Contrast method L	The ensemble model under the black-box attack
Query-based black-box attack	Attacker E	The attacker lacks any knowledge of the model collection or gradients and can only query the model probability vector to implement a black-box attack.		Contrast method M	The single model under the black-box attack

as dashed curves. Compared to the baseline models, we can observe that the ensemble model is highly vulnerable to white-box attacks, even worse than the single models. The random smoothing method improves the robustness of a single model, and the ensemble-smoothed model further improves the robustness and addresses the vulnerability of the ensemble under white-box attacks. Among all attack methods, attacker *B* has the worst attack performance, indicating that protecting the model from frequent access to gradient information at each iteration is crucial for SEM robustness. Attacker D, who has partial knowledge of the model collection but ensemble attributes in each iteration, can achieve transfer attacks through the ensemble and achieves similar robustness performance (even better than PGD) compared to attacker A. However, comparing the performance of attackers C and D, the SEM does not improve the attack transferability effect as a regularization method. This reveals the importance of protecting the model collection for SEM robustness. When the attacker has a higher transferability attack algorithm (for the MIM and PGD), the benefits of transferability are only for attacker *D* and are no longer attained by SEM. For the ensemble smoothed model (F curves) that has the SOTA performance between the contrasting baseline methods, the best attack performance cannot easily exceed the attack success rate associated with it.

Figure 5 shows the *ASR-vs.-distortion* curves for targeted transfer-based white-box attacks. When comparing different attack algorithms, the improved transferability of the PGD method does not significantly improve the attack performance under SEM. However, its robustness is significantly improved against the momentum-based attack, indicating that the randomness of the gradient at the model level has some impact on the confusion of the gradient direction. The variation in the attack knowledge of model collection between *A* and *C* does not significantly affect the robustness of SEM when against targeted attacks. However, contrary to the conclusion drawn from untargeted attacks, the robustness performance of SEM under *A* and *C* does not consistently exceed that of the ensemble smoothed or single smoothed model, demonstrating the lack of heterogeneity of the model in the gradient direction. However, as the detailed results in the second line of Figure 5 shown, the proposed method consistently demonstrates superior robustness under small perturbations. When comparing attackers *A, B, C,* and *D,* the weakest attack performance is exhibited by B (although this could be reversed when attacker D uses the PGD algorithm). Combined with the results of the untargeted attacks, we suggest that reducing the frequency of ensemble changes is critical for SEM when the model collection and ensemble attributes can be obtained by an attacker.

4.4.2. Query-based black-box analysis

The results of an untargeted source-based black-box attack are depicted in Figure 6A. The ensemble model exhibits weaker robustness to both NES and SPSA attacks compared to the single model, highlighting the vulnerability of the ensemble model to black-box attacks. Both the SPSA and NES approaches assume

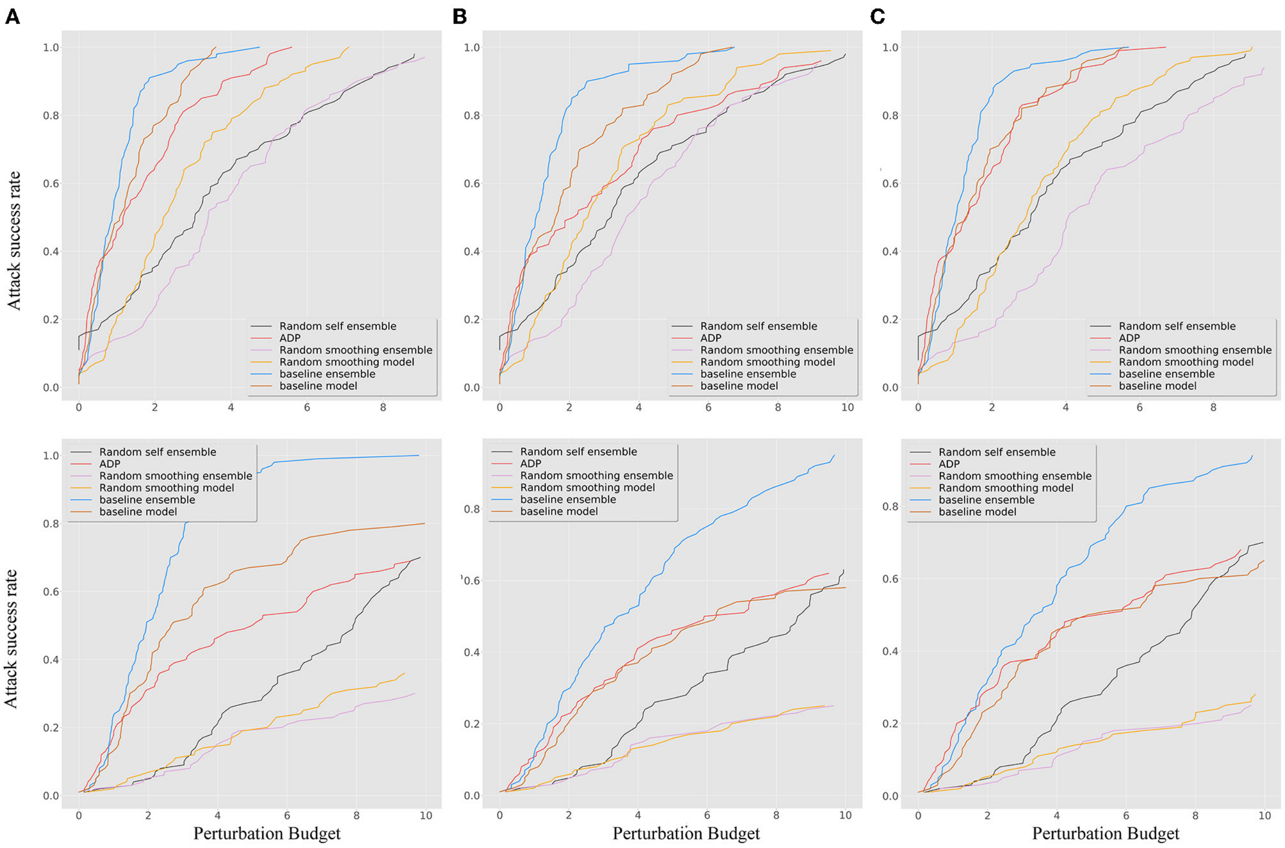


FIGURE 3
The *ASR-vs.-distortion* curves for different ensemble baseline methods under untargeted (first line) and targeted (second line) white-box attacks: **(A)** BIM; **(B)** MIM; and **(C)** PGD.

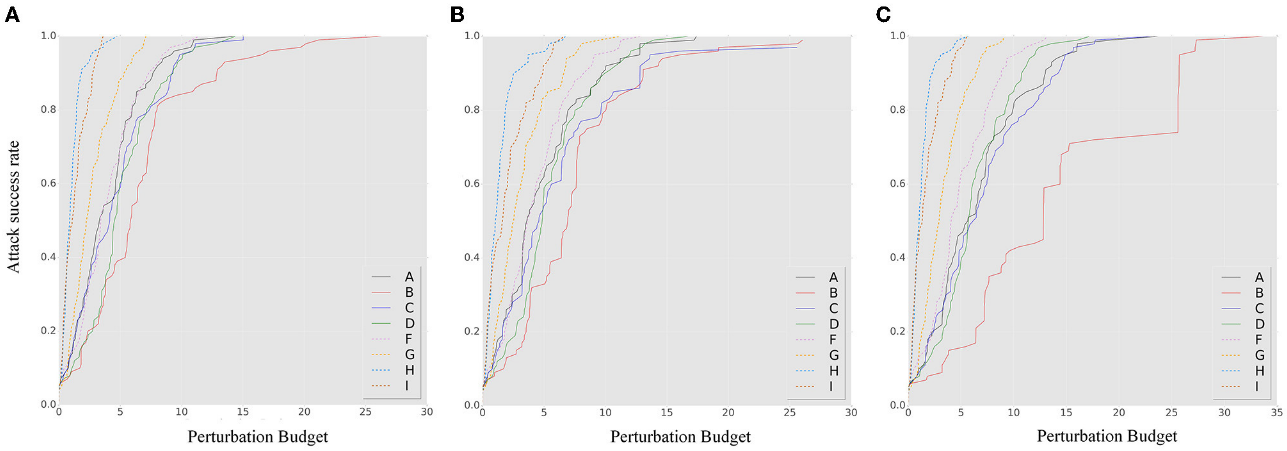


FIGURE 4
The *ASR-vs.-distortion* curves for untargeted transfer-based white-box attacks: **(A)** BIM, **(B)** MIM, **(C)** PGD. The A–D solid lines show the *ASR-vs.-distortion* curves under different attack capabilities while the dashed lines F–I show the curves under the contrast method. Compared to the two curves, the stochastic ensemble has better robustness even under the strongest adversary.

that the gradient direction of adversarial samples follows a certain probability distribution. This assumption is based on randomly sampling the gradient direction under a probability distribution, with the step size controlled by the loss value. The evaluation of the SEM under this expectation hypothesis is essentially a measure of the overlap between the gradient direction and the assumed distribution direction under the probability. In the experiment, the SEM does not demonstrate superior untargeted black-box defense

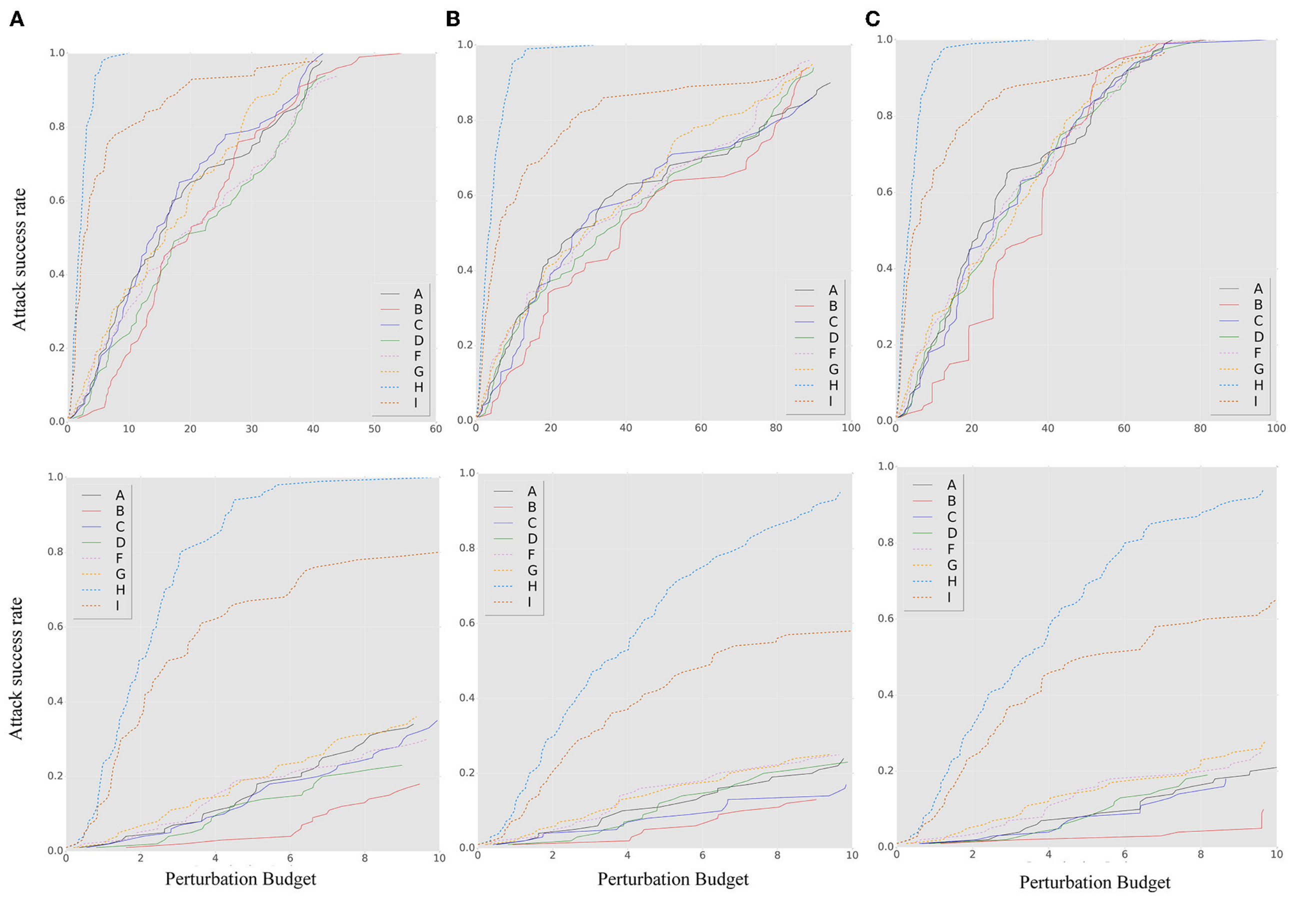


FIGURE 5
The *ASR-vs.-distortion* curves for targeted transfer-based white-box attacks: **(A)** BIM, **(B)** MIM, and **(C)** PGD. The solid lines A–D show the result under different attack capabilities while the dashed lines F–I show the curves under the contrast method. The second line shows the corresponding detail result with a small perturbation for clarity. From the detailed result, it can be concluded that the SEM exhibits superior robustness performance under conditions of small perturbation.

effectiveness compared to the smoothed ensemble, suggesting that the SEM based on different smoothing parameters may be more susceptible to high variance noise expectations (set δ as 1 for contrast method). We believe that this characteristic can be attributed to the high ensemble probability of an unsmoothed model or a smoothed model with low variance. As a result, the defensive effectiveness of SEM is not as impressive as that of the ensemble-smoothed model in terms of probability. This result highlights the influence of the smoothing model collection on the attack performance with respect to the smoothing parameter distribution.

In comparison, the results for the targeted source-based black-box attacks that show a decrease in overall accuracy are shown in Figure 6B. Nevertheless, the same conclusion regarding robustness can be drawn. The sensitivity of the model to specific noise distributions was analyzed through experiments with black-box attacks, and it was found that the smoothing model resulted in improved defense performance against adversarial samples based on specific noise distribution assumptions. However, the model's susceptibility to noise with varying parameters under different smoothing parameters limits its defense capabilities. Such noise assumptions are independent of the true gradient information of the model and rely primarily on changes in the model output and the number of queries. Improvements in the selection of smoothing parameters for the ensemble strategy are needed to further enhance the defensive capabilities.

4.5. Robustness analysis based on the stochastic ensemble strategy

This section examines the effect of ensemble quantity and heterogeneity on the robustness of the proposed method. Specifically, we compare ensembles with quantities of 1, 2, and 3 to those with quantities of 6, 7, and 8 (multi_ensemble). In addition, we compare a stochastic ensemble consisting of a single-architecture CNN with different smoothing parameters. To ensure comparable prediction accuracies with our method, we choose the WRN (Zagoruyko and and Komodakis, 2016b) as the single-architecture neural network (single_architecture). To expand the stochastic ensemble model collection space and introduce model gradient variations, we smooth the WRN using seven different smoothing parameters (0.12, 0.15, 0.25, 0.5, 0.75,

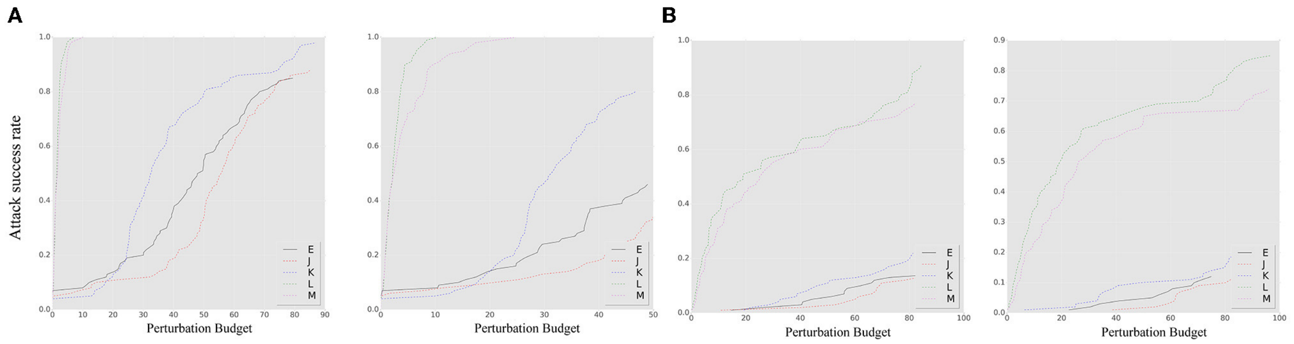


FIGURE 6
The *ASR-vs.-distortion* curves for source-based black-box attacks: **(A)** untargeted attack; **(B)** targeted attack. The left side of each attack target represents the NES, while the right side represents the SPSA. The solid line E shows the result of the SEM under a source-based attack. The dashed lines J–M represent the curve under the contrast method.

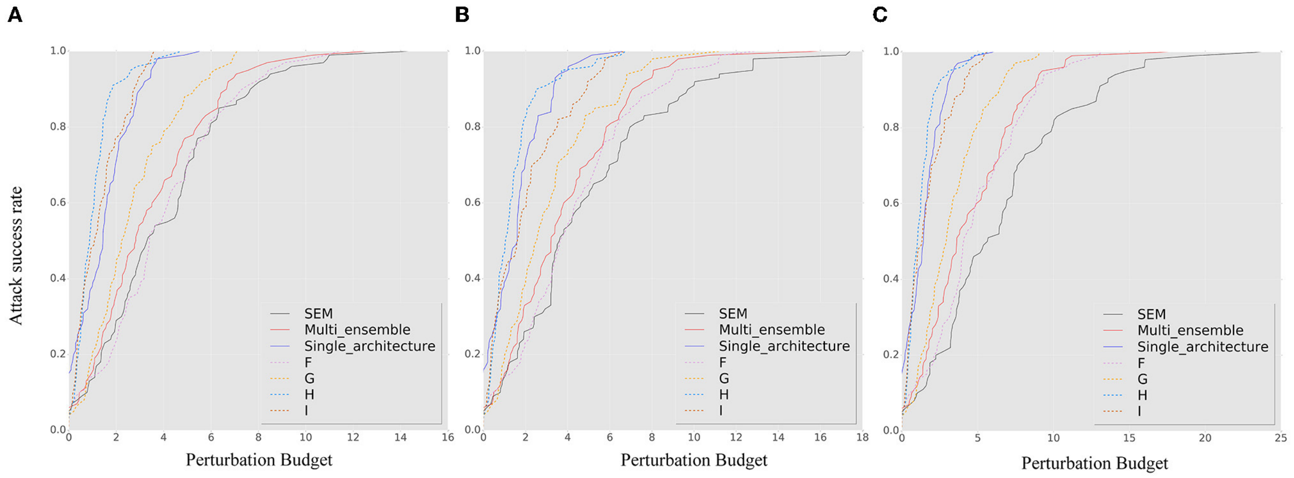


FIGURE 7
The *ASR-vs.-distortion* curves for untargeted white-box attacks under different attack methods and ensemble strategies: **(A)** BIM, **(B)** MIM, and **(C)** PGD. The solid lines represent the *ASR-vs.-distortion* curves under different ensemble strategies, while the dashed lines represent the same curves under the contrast method for comparison.

1.0, and 1.25) under Gaussian noise *via* stability training (Li B. et al., 2019), semi-supervised learning (Carmon et al., 2019), and pre-training (Hendrycks et al., 2019). The resulting stochastic ensemble, consisting of a single-architecture CNN, shows heterogeneity in its smoothing attributes.

Figures 7, 8 show the results of our robustness evaluation using different ensemble strategies. The negative impact of ensemble quantity on robustness is evident, as shown by the red solid line. As explained in Section 3.3, a larger ensemble quantity leads to reduced gradient differences and increased transferability of adversarial samples across ensemble iterations. The blue solid line in Figure 7 indicates that architectural heterogeneity has a greater impact on the adversarial robustness of the SEM. When there are no architectural differences between the ensemble models, even in the random smoothing case, the SEM can actually increase vulnerability to adversarial samples.

Figure 8 confirms that an SEM without architectural heterogeneity is even more vulnerable than an ensemble model. Viewing the ensemble strategy of SEM as a form of dropout operation (Baldi and Sadowski, 2013), we observe that when the ensemble quantity is large and there is insufficient architectural diversity, the SEM method becomes a regularization technique that conversely enhances the capability of adversarial samples, especially under targeted attack.

5. Conclusion

This study proposes a dynamic defense method for the generalized robustness of deep neural networks based on random smoothing. This dynamic nature based on the ensemble system is a change from the perspective of the existing random method from the model level to the system level. The ensemble attributes are considered as the changeable factor and dynamically adjusted during the inference prediction phase. The proposed method

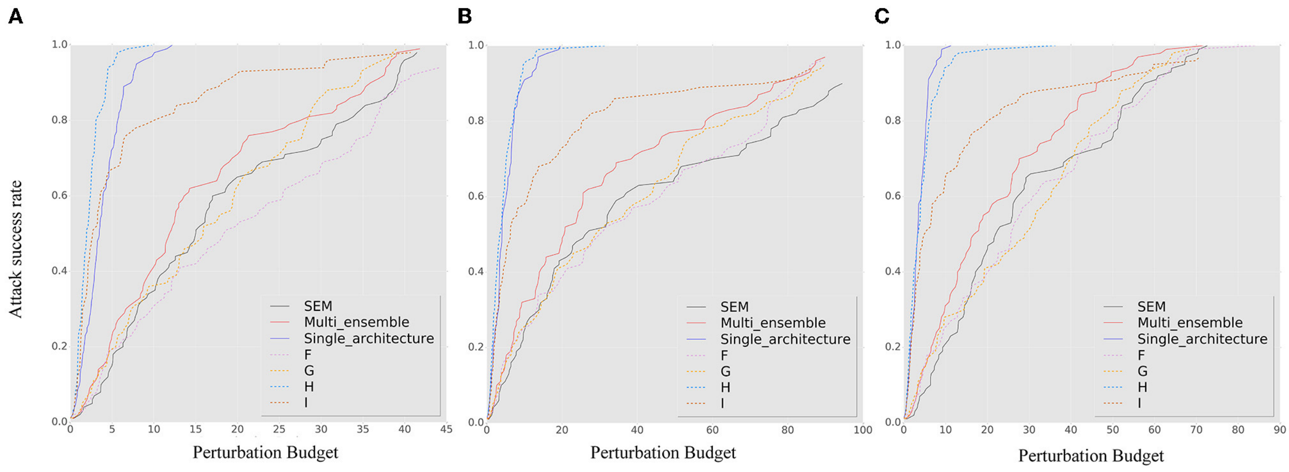


FIGURE 8
The *ASR-vs.-distortion* curves for white-box targeted attacks under different attack methods and ensemble strategies: **(A)** BIM, **(B)** MIM, and **(C)** PGD. The solid lines represent the *ASR-vs.-distortion* curves under different ensemble strategies, while the dashed lines represent the same curve under the contrast method.

has the characteristics of diversity, randomness, and dynamics to achieve the probabilistic attribute dynamic defense for adversarial robustness without damaging the accuracy of clean samples. Through an optimal search of perturbation values under different attack capabilities, attack methods, and attack targets according to the degree of the real-time ability of an attacker to obtain knowledge of the model collection and gradients, a comprehensive evaluation under CIFAR10 preliminarily demonstrates that when the image distortion is small, even the attacker with the highest attack capability cannot easily exceed the attack success rate associated with the ensemble smoothed model, especially under untargeted attacks.

The robustness of our proposed method relies heavily on the heterogeneity and confidentiality of the model collection. Through experimental setups under different attack scenarios, this study also finds that the proposed SEM can achieve better robustness by limiting the ability of the adversary. Therefore, based on these findings, future studies will be conducted (1) to further improve the robustness against white-box attacks, adaptive control of the ensemble changes based on attack detection is a crucial research direction; (2) under the query-based black-box analysis, the smooth parameter selection probability of the ensemble strategy is a crucial optimization direction for this study; (3) for practical applications, both the number of parameters of the model and the forward efficiency of the ensemble prediction should be considered. In this study, the robustness is evaluated on the CIFAR10 dataset, but there are practical application problems because of the large training cost. Therefore, the light weight of the ensemble model is an important research direction.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

RQ: conceptualization, validation, software, and writing—original draft. LW: methodology, resources, and supervision. XD: funding acquisition, investigation, and supervision. PX: supervision. XC: project administration and funding acquisition. BY: writing—reviewing and editing. All authors contributed to the article and approved the submitted version.

Funding

This document is the result of the research project funded by the Communication Document of Henan Provincial Committee of China under Grant 44 by XD.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Akhtar, N., and Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: a survey. *IEEE Access* 6, 14410–14430. doi: 10.1109/ACCESS.2018.2807385

Athalye, A., Engstrom, L., Ilyas, A., and Kwok, K. (2018). "Synthesizing robust adversarial examples," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, eds J.G. Dy and A. Krause (Stockholmsmässan, Stockholm: ICML), 284–293.

Attiah, A., Chatterjee, M., and Zou, C. C. (2018). "A game theoretic approach to model cyber attack and defense strategies," in *2018 IEEE International Conference on Communications (ICC)* (Piscataway, NJ: IEEE), 1–7.

Baldi, P., and Sadowski, P. J. (2013). "Understanding dropout," in *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*. Lake Tahoe, NV, United States, 2814–2822.

Bojarski, M., Testa, D. D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., at al. (2014). End to end learning for self-driving cars. *arXiv [Preprint].* arXiv: 1604.07316. Available online at: https://arxiv.org/abs/1604.07316

Carlini, N., and Wagner, D. (2017). "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy* (sp). (Piscataway, NJ: IEEE), 39–57.

Carmon, Y., Raghunathan, A., Schmidt, L., Duchi, J. C., and Liang, P. S. (2019). "Unlabeled data improves adversarial robustness," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, ed H.M. Wallach, Vancouver, BC, Canada, 11190–11201.

Cohen, J. M., Rosenfeld, E., and Kolter, J. Z. (2019). "Certified Adversarial Robustness via Randomized Smoothing," in *Proceedings of the 36th International Conference on Machine Learning*, eds K. Chaudhuri and R. Salakhutdinov (Long Beach, CA, USA: ICML), 1310–1320.

Croce, F., Gowal, S., Brunner, T., Shelhammer, E., Hein, M., and Cemgil, T. (2022). Evaluating the adversarial robustness of adaptive test-time defenses. *arXiv [Preprint].* arXiv: 2202.13711. Available online at: https://arxiv.org/abs/2202.13711

Dong, Y., Fu, Q. A., Yang, X., Pang, T., Su, H., Xiao, Z. (2019). "Benchmarking Adversarial Robustness," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT, USA: CVPR).

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., and Li, J. (2018). "Boosting Adversarial Attacks With Momentum," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018* (Salt Lake City, UT, USA: CVPR), 9185–9193.

Fischetti, M., and Jo, J. (2017). Deep neural networks as 0-1 mixed integer linear programs: A feasibility study. *arXiv [Preprint].* arXiv: 1712.06174. Available online at: http://arxiv.org/abs/1712.06174

Gao, H., Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). "Densely Connected Convolutional Networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE Conference on Computer Vision and Pattern Recognition* (Piscataway, NJ: IEEE), 4700–4708.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). "Explaining and harnessing adversarial examples," in *3rd International Conference on Learning Representations*, eds Y. Bengio and Y. LeCun (San Diego, CA, USA: ICLR).

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Piscataway, NJ: IEEE), 770–778.

He, W., Wei, J., Chen, X., Carlini, N., and Song, D. (2017). "Adversarial example defense: Ensembles of weak defenses are not strong," in *11th USENIX Workshop on Offensive Technologies (WOOT 17*).

He, Y., Meng, G., Chen, K., Hu, X., and He, J. (2020). *Towards Security Threats of Deep Learning Systems: A Survey. IEEE Transactions on Software Engineering.* Piscataway, NJ: IEEE.

Hendrycks, D., Lee, K., and Mazeika, M. (2019). "Using pre-training can improve model robustness and uncertainty," in *Proceedings of the 36th International Conference on Machine Learning*, eds K. Chaudhuri and R. Salakhutdinov (Long Beach, CA, USA: ICML), 2712–2721.

Hu, S., Shen, Y., Wang, S., and Lei, B. (2020). "Brain MR to PET synthesis via bidirectional generative adversarial network," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference*, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23 (Cham: Springer), 698–707.

Hung, K., and Fithian, W. (2016). Rank verification for exponential families. *arXiv [Preprint].* arXiv: 1610.03944. Available online at: http://arxiv.org/abs/1610.03944

Ilyas, A., Engstrom, L., Athalye, A., and Lin, J. (2018). "Black-box Adversarial Attacks with Limited Queries and Information," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, eds J.G. Dy and A. Krause (Stockholm, Sweden: ICML), 2142–2151.

Jajodia, S., Ghosh, A. K., Swarup, V., Wang, C., and Wang, X. S. (2011) *Moving Target Defense - Creating Asymmetric Uncertainty for Cyber Threats*. Cham: Springer

Krizhevsky, A., and Hinton, G. (2009) *Learning Multiple Layers of Features from Tiny Images' (2009).* Citeseer.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90. doi: 10.1145/3065386

Kurakin, A., Goodfellow, I., Bengio, S., Dong, Y., and Liao, F. (2018). "Adversarial attacks and defences competition," in *The NIPS'17 Competition: Building Intelligent Systems* (Cham: Springer), 195–231.

Kurakin, A., Goodfellow, I. J., and Bengio, S. (2016). *Adversarial Examples in the Physical World.* London: Chapman and Hall.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, (Long Beach, CA), 6402–6413.

Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. (2019). "Certified robustness to adversarial examples with differential privacy," in *2019 IEEE Symposium on Security and Privacy*, SP 2019 (San Francisco, CA, USA: IEEE), 656–672.

Li, B., Chen, C., Wang, W., and Carin, L. (2019). "Certified adversarial robustness with additive noise," in *Annual Conference on Neural Information Processing Systems 2019*, Vancouver, BC, Canada, 9459–9469.

Li, Y., Li, L., Wang, L., Zhang, T., and Gong, B. (2019). "NATTACK: learning the distributions of adversarial examples for an improved black-box attack on deep neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, eds K. Chaudhuri and R. Salakhutdinov (Long Beach, California, USA: ICML), 3866–3876.

Liu, C., Feng, Y., Wang, R., and Dong, B. (2020). *Enhancing Certified Robustness of Smoothed Classifiers via Weighted Model Ensembling*, CoRR abs/2005.09363. Available online at: https://arxiv.org/abs/2005.09363

Liu, X., Cheng, M., Zhang, H., and Hsieh, C. J. (2018). "Towards robust neural networks via random self-ensemble," in *Proceedings of the European Conference on Computer Vision (ECCV),* 369–385.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2018). "Towards deep learning models resistant to adversarial attacks," in *6th International Conference on Learning Representations* (Vancouver, BC, Canada: ICLR).

Pang, T., Xu, K., Du, C., Chen, N., and Zhu, J. (2019). "Improving adversarial robustness via promoting ensemble diversity," in *Proceedings of the 36th International Conference on Machine Learning*, eds K. Chaudhuri and R. Salakhutdinov (Long Beach, California, USA: ICML), 4970–4979.

Papernot, N., McDaniel, P., Wu, X., Jha, S., and Swami, A. (2016). "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE Symposium on Security and Privacy (SP)* (Piscataway, NJ: IEEE), 582–597.

Parkhi, O. M., Vedaldi, A., Zisserman, A. (2015). "Deep face recognition." in: *Proceedings of the British Machine Vision Conference 2015, BMVC 2015*, (Swansea: BMVA Press), 41–14112. doi: 10.5244/C.29.41

Perez, L., and Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv [Preprint].* arXiv: 1712.04621. Available online at: https://arxiv.org/abs/1712.04621

Ren, S., He, K., Girshick, R., and Sun, J. (2016). Faster r-cnn: Towards real-time object detection with region proposal networks.n *IEEE Trans. Pattern Anal. Machine Int.* 39, 1137–1149. doi: 10.1109/TPAMI.2016.2577031

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L. C. (2018). "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Piscataway, NJ: IEEE), 4510–4520.

Simonyan, K., and Zisserman, A. (2014). "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015* (San Diego, CA). Available online at: http://arxiv.org/abs/1409.1556

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* Piscataway (NJ: IEEE), 2818–2826.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., and Erhan, D. (2014). "Intriguing properties of neural networks," in *2nd International Conference on Learning Representations*, eds Y. Bengio and Y. LeCun Banff (Toronto: ICLR).

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D. (2015). "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Piscataway, NJ: IEEE), 1–9.

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. (2018). "Ensemble adversarial training: attacks and defenses," in *6th International Conference on Learning Representations* (Vancouver, BC, Canada: ICLR).

Uesato, J., O'donoghue, B., Kohli, P., and Oord, A. (2018). "Adversarial risk and the dangers of evaluating against weak attacks," in *Proceedings of the 35th International*

Conference on Machine Learning, eds J.G. Dy and A. Krause (Stockholm: ICML), 5032–5041.

Wenzel, F., Snoek, J., Tran, D., and Jenatton, R. (2020). Hyperparameter ensembles for robustness and uncertainty quantification. *Adv. Neural Inf. Proc. Syst.* 33, 6514–6527.

Wu, J. (2020). *Cyberspace Mimic Defense - Generalized Robust Control and Endogenous Security*. Cham: Springer.

Wu, Z., Chen, X., Yang, Z., and Du, X. (2019). Reducing security risks of suspicious data and codes through a novel dynamic defense model. *IEEE Trans. Inf. Forensics Secur.* 14, 2427–2440. doi: 10.1109/TIFS.2019.2901798

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Piscataway, NJ: IEEE), 1492–1500.

You, S., Lei, B., Wang, S., Chui, C. K., Cheung, A. C., Liu, Y., et al. (2022). *Fine Perceptive GANs for Brain MR Image Super-Resolution in Wavelet Domain. IEEE Transactions on Neural Networks and Learning Systems*. Piscataway, NJ: IEEE.

Zagoruyko, S., and Komodakis, N. (2016a). *Wide Residual Networks*.

Zagoruyko, S., and Komodakis, N. (2016b). "Wide residual networks," in *Proceedings of the British Machine Vision Conference 2016*, eds R. C. Wilson, E. R. Hancock, and W. A. P. Smith (New York, NY: BMVA Press). Available online at: http://www.bmva.org/bmvc/2016/papers/paper087/index.html

Zhang, R., Xia, H., Hu, C., Zhang, C., Liu, C., and Xiao, F. (2022). generating adversarial examples with shadow model. *IEEE Trans. Ind. Inf.* 18, 6283–6289. doi: 10.1109/TII.2021.3139902

TYPE Original Research
PUBLISHED 14 August 2023
DOI 10.3389/fnbot.2023.1193011

Check for updates

OPEN ACCESS

EDITED BY
Long Jin,
Lanzhou University, China

REVIEWED BY
Varun Dutt,
Indian Institute of Technology Mandi, India
Hang Su,
Fondazione Politecnico di Milano, Italy

*CORRESPONDENCE
Qing He
✉ qhe@gzu.edu.cn

RECEIVED 24 March 2023
ACCEPTED 21 July 2023
PUBLISHED 14 August 2023

CITATION
Li Y, He Q and Zhang D (2023) Dual graph convolutional networks integrating affective knowledge and position information for aspect sentiment triplet extraction.
Front. Neurorobot. 17:1193011.
doi: 10.3389/fnbot.2023.1193011

Dual graph convolutional networks integrating affective knowledge and position information for aspect sentiment triplet extraction

Yanbo Li, Qing He* and Damin Zhang

College of Big Data and Information Engineering, Guizhou University, Guiyang, China

Aspect Sentiment Triplet Extraction (ASTE) is a challenging task in natural language processing (NLP) that aims to extract triplets from comments. Each triplet comprises an aspect term, an opinion term, and the sentiment polarity of the aspect term. The neural network model developed for this task can enable robots to effectively identify and extract the most meaningful and relevant information from comment sentences, ultimately leading to better products and services for consumers. Most existing end-to-end models focus solely on learning the interactions between the three elements in a triplet and contextual words, ignoring the rich affective knowledge information contained in each word and paying insufficient attention to the relationships between multiple triplets in the same sentence. To address this gap, this study proposes a novel end-to-end model called the Dual Graph Convolutional Networks Integrating Affective Knowledge and Position Information (DGCNAP). This model jointly considers both the contextual features and the affective knowledge information by introducing the affective knowledge from SenticNet into the dependency graph construction of two parallel channels. In addition, a novel multi-target position-aware function is added to the graph convolutional network (GCN) to reduce the impact of noise information and capture the relationships between potential triplets in the same sentence by assigning greater positional weights to words that are in proximity to aspect or opinion terms. The experiment results on the ASTE-Data-V2 datasets demonstrate that our model outperforms other state-of-the-art models significantly, where the F1 scores on 14res, 14lap, 15res, and 16res are 70.72, 57.57, 61.19, and 69.58.

KEYWORDS
aspect-based sentiment analysis, aspect sentiment triplet extraction, affective knowledge, position-aware function, graph convolutional network (GCN)

1. Introduction

In recent years, significant advancements in deep learning have been attributed to the development of more efficient algorithms, advancements in hardware capabilities, and the availability of extensive datasets. These progressions have paved the way for the emergence of diverse types of dynamic neural networks (DNN) tailored to address specific challenges across various domains. For instance, deep learning has been instrumental in surface defect recognition in the realm of computer vision (Shi et al., 2023), Artificial Intelligence (AI) systems based on deep learning algorithms can effectively detect and analyze arc faults in electrical systems (Tian et al., 2023) and recurrent neural networks (RNN) are designed to capture temporal dependencies and sequential patterns, thus making them well suited for tasks involving gesture recognition and classification. Moreover, the utilization of graph structures for learning purposes has demonstrated tremendous potential in various fields. For example, in the domain of blockchain technology, graph structure learning methods have been employed to enhance the analysis of transaction networks and identify the characteristics of the transaction (Wang et al., 2023). Additionally, improved graph structure learning methods (Liu et al., 2023) based on the foundational graph neural network (GNN) have been proposed in order to further enhance the capabilities of graph-based learning.

In the field of natural language processing (NLP), comments of consumers serve as a valuable resource for gathering information that can aid in enhancing the performance of robots and their associated products or services. With the proliferation of social media communities, the availability of consumer-generated content has expanded significantly, presenting an opportunity to leverage this data for insights and improvements. By employing methods designed for text information, robots can significantly enhance their ability to understand the intent and meaning behind a comment of consumer. These methods enable robots to extract the most valuable information from user input, leading to more accurate and meaningful interactions. Aspect Sentiment Triplet Extraction (ASTE) (Peng et al., 2020) is concerned with identifying the triplets from a given comment. Each triplet includes an aspect term, corresponding opinion term, and the sentiment polarity of this aspect term. For instance, in Figure 1, this comment from restaurant domain comprises two triplets: (*menu, limited, negative*) and (*dishes, excellent, positive*). Aspect sentiment triplet extraction plays a crucial role in enabling a more fine-grained understanding of text by capturing sentiments toward specific aspects or features. This capability facilitates context-aware analysis, supports decision-making processes, analyzes customer feedback, and aids in brand monitoring and reputation management.

Aspect Sentiment Triplet Extraction (ASTE) is a fine-grained task of Aspect-based Sentiment Analysis (ABSA) (Pontiki et al., 2014). ABSA aims to extract aspect terms and identify the corresponding sentiment polarity from a given sentence. It typically includes subtasks such as Aspect Terms Extraction (ATE) (Yin et al., 2016; Xin et al., 2018; Wu et al., 2020b), Opinion Terms Extraction (OTE) (Jebbara and Cimiano, 2017; Jordhy et al., 2019; Li et al., 2019), and Aspect-based Sentiment Classification (ASC) (Tang et al., 2016; Ma et al., 2017; He et al., 2018). ASTE is the combination of these subtasks and initially proposed in the study by Peng et al. (2020) with a two-stage pipeline approach. This method predicts all aspect terms, opinion terms, and sentiment polarities in the first stage. In the second stage, aspect terms are paired with their corresponding opinion terms to obtain triplets. However, this approach is susceptible to error propagation. To overcome this limitation, Xu et al. (2020) propose a position-aware tagging scheme and develop a union model that uses sequence labeling to extract triplets. This method is the first end-to-end model in the ASTE task. Similarly, Wu et al. (2020a) present a grid tagging scheme named GTS that uses a unified grid markup task to extract triplets in an end-to-end manner.

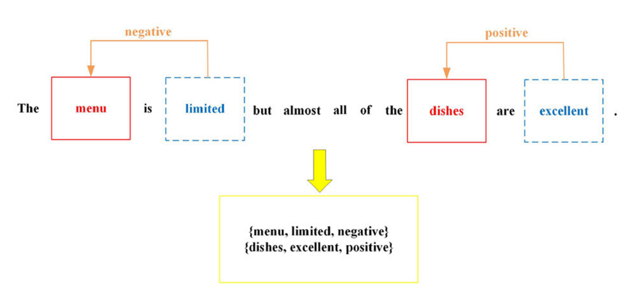


FIGURE 1
An example of ASTE. The aspect terms are highlighted in red. The terms in blue are opinion terms and the origin words that denote their sentiment polarity. All triplets are shown in the yellow box.

During sentiment analysis, it is observed that every word in a sentence possesses a unique emotional intensity. For instance, while words such as "likable" and "charming" both convey a positive sentiment polarity, their degrees of positivity differ. However, it has been noted that current networks relying on graph convolutional network tend to utilize solely syntactic dependencies for graph construction, thereby ignoring the commonsense knowledge information (Erik et al., 2009) associated with each word. Furthermore, such models typically overlook the relationships between multiple triplets present in the same sentence.

To overcome the aforementioned limitations of existing models, this study presents a novel approach that takes into account both affective knowledge information and the implicit relationship between different potential triplets in the same sentence. The proposed method employs a part-of-speech (POS) based approach to identify potential aspect terms and opinion terms within sentences, then formulates a fresh approach for generating an adjacency matrix, which fuses the affective score of each word from SenticNet (Ma et al., 2018) with the syntax dependency in two parallel modules, leading to the generation of a potential aspect terms enhanced adjacency matrix and a potential opinion terms enhanced adjacency matrix. These adjacency matrices are, then, input into a graph convolutional network (GCN) (Kipf and Welling, 2016) to extract features separately. GCN is a neural network architecture that has the ability to extract both contextual and syntactic representations from the adjacency matrix by aggregating the features of neighboring nodes. Additionally, this study utilizes a multi-target position-aware function in each GCN module, which assigns different weights to all words based on the position of potential aspect words or opinion words. This facilitates interaction between different potential triplets in a sentence and reduces interference from other words on triplet extraction. Finally, the hidden representations produced by the encoder layer, and two GCN modules are used *via* GTS for triplet extraction.

The main contributions of our study can be summarized as follows:

- We propose an innovative Dual Graph Convolutional Networks Integrating Affective Knowledge and Position

Information (DGCNAP) for the ASTE task in an end-to-end manner.

- We conceive a novel method to introduce affective knowledge information into the adjacency matrix generated by sentences in the ASTE task.
- We design a multi-target position-aware function in the GCN layer to reduce interference and capture the associations between different potential triplets in the same sentence.
- Our experimental results on four benchmark datasets demonstrate the effectiveness of our model in the ASTE task.

2. Related work

Unlike traditional sentiment analysis that aims to identify the sentiment polarity of the whole document or sentence, ABSA aims to predict sentiment polarity of specific aspect terms. In recent research, most models use attention mechanisms. Wu et al. (2022) proposed a phrase dependency graph attention network to aggregate directed dependency edges and phrase information. Liang et al. (2022) adopted a graph convolutional network based on affective knowledge to leverage the affective dependencies of the sentence; thus, both the dependencies of contextual words and aspect words and the affective information between opinion words and the aspect are considered.

To establish a comprehensive solution for ABSA, ASTE aims to complete multiple subtasks of ABSA simultaneously. In the ASTE task, existing methods can be divided into two types: pipeline methods and end-to-end methods. Peng et al. (2020) are the first to propose a complete solution for the ASTE task, employing a two-stage pipeline approach. However, models constructed using this pipeline approach are rather simple and are easily affected by error propagation. To avoid this problem, end-to-end models have been proposed and can be summarized as follows. Xu et al. (2020) first developed an end-to-end method named position-aware tagging scheme. Similarly, Wu et al. (2020a) proposed grid tagging scheme to extract triplets simultaneously. Considering ASTE is the combination of all basic tasks of ABSA, Chen et al. (2022) proposed an end-to-end approach which decomposes ASTE into three subtasks, namely, target tagging, opinion tagging, and sentiment tagging. Chen et al. (2021) proposed a novel method which transforms ASTE task into a multi-turn machine reading comprehension task and propose a bidirectional MRC framework to address this challenge. Another end-to-end method (Dai et al., 2022) proposed a sentiment-dependence detector based on a dual-table structure that starts from two directions, aspect-to-opinion and opinion-to-aspect, to generate two sentiment-dependence tables dominated by two types of information. Shi et al. (2022) proposed an interactive attention mechanism to jointly consider both the contextual features and the syntactic dependencies in an iterative interaction manner. Previous tag-based joint extraction methods have been observed to struggle with effectively handling one-to-many and many-to-one relationships between aspect terms and opinion terms within sentences. This limitation has motivated researchers to explore alternative approaches, such as those that operate at the span level rather than relying on tagging schemes. A tagging-free approach (Mukherjee et al., 2021) is proposed to capture the span-level semantics while predicting the sentiment between an aspect-opinion pair. Li et al. (2022) proposed a span-sharing joint extraction framework to extract aspect terms and their corresponding opinion terms simultaneously in the last step, thereby avoiding error propagation. Hu et al. (2023) used a span GCN for syntactic constituency parsing tree and a relational GCN (R-GCN) for commonsense knowledge graph to build an end-to-end model for the ASTE task. Moreover, a double-embedding mechanism-character-level and word-vector embeddings are introduced for the first time. Zhang et al. (2022) propose a dual convolutional neural network with a span-based tagging scheme to extract multiple entities directly under the supervision of span boundary detection.

3. Approach

Existing models have achieved good performance on the ASTE task. However, a significant number of these methods disregard the abundant affective knowledge present in individual words of a sentence, as well as the interdependence of various triplets. To address this limitation, we introduce affective knowledge information in our framework while constructing the dependency graph. Additionally, we utilize a multi-target position-aware function to capture the interdependence of multiple triplets in the same sentence, and it can also mitigate the adverse effects of noisy words.

This section commences with a definition of the ASTE task followed by an elaborate elucidation of our proposed methodology, Dual Graph Convolutional Networks Integrating Affective Knowledge and Position Information (DGCNAP), for the ASTE task.

3.1. Definition of ASTE

Given an n-word sentence $\mathcal{S} = \{w_1, w_2, ..., w_n\}$, the ASTE task aims at identifying all sentiment triplet sets $\mathcal{T} = \{at, ot, s\}$, where "at" denotes the aspect term, "ot" denotes the opinion term, "s" denotes the sentiment of the aspect term in this set, and $s \in \{positive, negative, neutral\}$.

3.2. The DGCNAP framework

The overall architecture of DGCNAP model is shown in Figure 2. The model takes two parallel channels to joint potential aspect term and potential opinion term enhanced features extraction, leveraging affective knowledge, graph convolutional network, and multi-target position-aware function to improve accuracy and capture the complex relationships between aspect and opinion terms in sentences.

3.3. Embedding and encoding layers

In this study, we employ two types of encoders to learn hidden representations: the first is the Bi-directional Long Short-Term Memory (Bi-LSTM) (Hochreiter and Schmidhuber, 1997) network

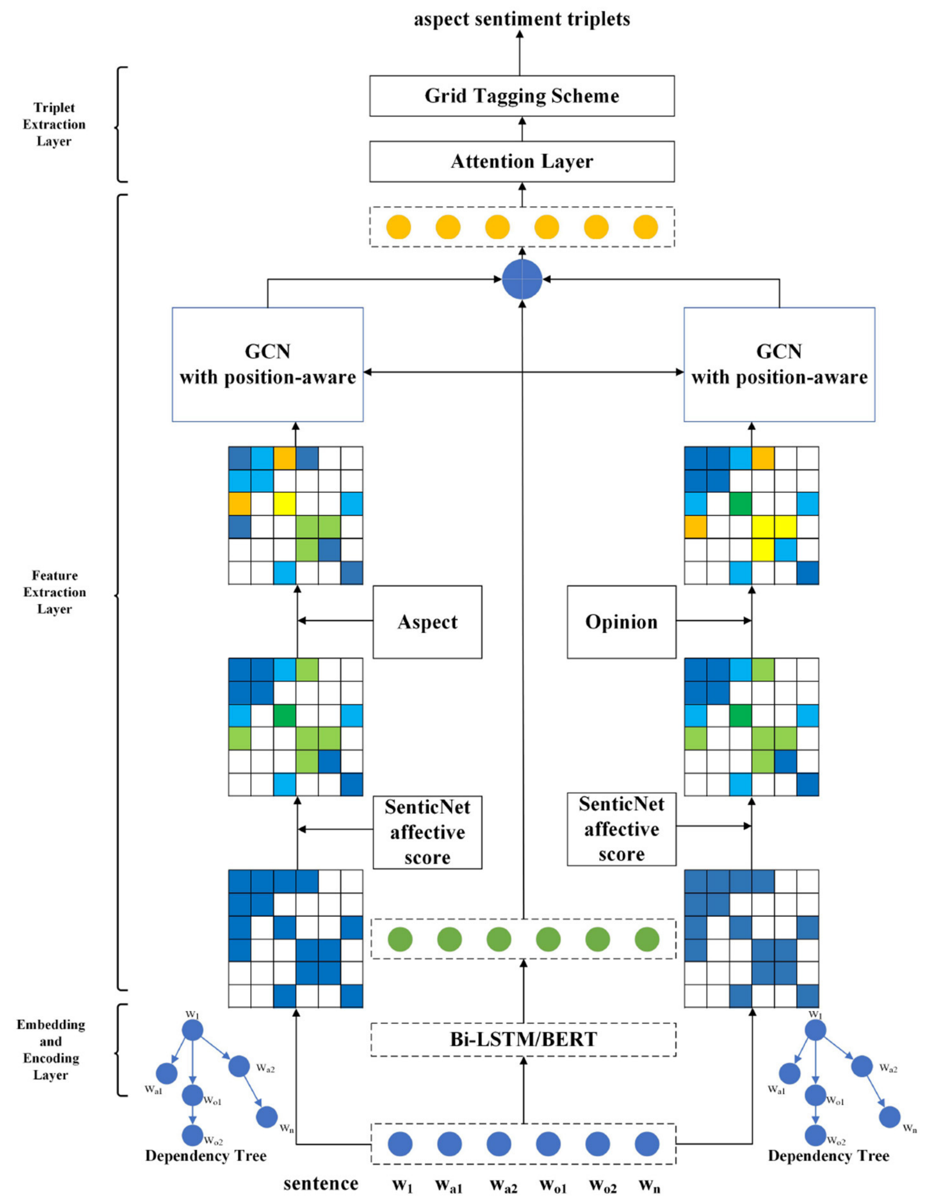


FIGURE 2
Architecture of DGCNAP.

and the second is the pre-trained language model BERT (Devlin et al., 2019).

For the Bi-LSTM-based encoder, we utilize double embedding to obtain the initial word representation and capture the contextual meaning of words in a specific domain. The specific-domain embedding was pre-trained based on the skip-gram model, where each word is represented as a bag of character n-grams. A vector representation is associated with each character n-gram; words are represented as the sum of these representations. We concatenate the 300-dimension general-domain embedding $E_w \in \mathbb{R}^{n\times d_w}$ and the 100-dimension specific-domain embedding $E_s \in \mathbb{R}^{n\times d_s}$ to form the final word representation $E \in \mathbb{R}^{n\times(d_w+d_s)}$, where d_w and d_s denote the dimensions of word embedding. After that, we input the embedding matrix into a Bi-LSTM to obtain the hidden contextual representations $H^c = \{h_1, h_2, ..., h_n\} \in \mathbb{R}^{n\times d_l}$ of the input sentence, where d_l denotes the hidden state dimension of Bi-LSTM:

$$H^c = Bi - LSTM(E) \qquad (1)$$

For the BERT-based encoder, we first add the [CLS] token at the beginning of the sentence and the [SEP] token at the end. Next, we feed the sequence into BERT for context encoding by converting it into a vector that sums its token embedding, segment embedding, and position embedding. Finally, we input the vector v into the transformer encoder (Vaswani et al., 2017), to obtain the hidden contextual representation $H^c = \{h_1, h_2, ..., h_n\} \in \mathbb{R}^{n\times d_l}$:

$$H^c = BERT(v) \qquad (2)$$

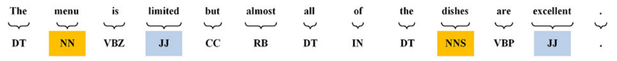


FIGURE 3
An example of part-of-speech tagging.

3.4. Generate enhanced graph

Part-of-speech (POS) is a linguistic concept that categorizes words based on their grammatical roles and syntactic functions within a sentence. Each word in a sentence is assigned a specific part-of-speech tag, which provides information about its linguistic characteristics and relationships with other words. As shown in Figure 3, the aspect terms "menu" and "dishes" are both annotated as nouns, and the opinion terms "limited" and "excellent" are both annotated as adjectives. In the proposed approach, nouns are considered as potential aspect terms, while adjectives are identified as potential opinion terms.

Dependency graph is a useful way to represent the grammatical relationships between words in a sentence. We use the dependency tree of each input sentence to construct a unidirectional dependency graph with self-loop. $D \in \mathbb{R}^{n\times n}$ denotes the adjacency matrix obtained from the graph:

$$D_{i,j} = \begin{cases} 1 & \text{if } w_i \text{ and } w_j \text{ contains dependency} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Because the parent node is also affected by the child node, $D_{j,i} = D_{i,j}$.

To incorporate affective knowledge into the construction of the dependency graph, we take the absolute value of the SenticNet affective score and use it as a weight for the corresponding edge in the adjacency matrix. By doing so, we can assign more weight to words with stronger sentiment intensity when computing the graph convolution operation, and our model can learn meaningful information from words containing emotionally intense, thereby contributing to increased accuracy in predicting sentiment polarity corresponding to aspect terms:

$$S_{i,j} = |SenticNet(w_i)| + |SenticNet(w_j)| \tag{4}$$

where $SenticNet(w_i) \in [-1, 1]$ denotes the SenticNet affective score of word w_i. When $SenticNet(w_i)$ approaches -1, the word conveys a strong negative sentiment. Conversely, as $SenticNet(w_i)$ approaches 1, the word expresses a strong positive sentiment. In cases where $SenticNet(w_i)$ is equal to 0, the word w_i is considered neutral or is not included in the SenticNet database. We exploit SenticNet 6, which contains 200,000 concepts. Some examples of SenticNet are shown in Table 1.

To enhance the sentiment dependencies that exist between potential aspect words and contextual words, as well as between potential opinion words and contextual words, we incorporate potential aspect word weights and potential opinion word weights as the target score into the generation of the adjacency matrix:

$$T^a_{i,j} = \begin{cases} 1 & \text{if } w_i \text{ or } w_j \text{ is a potential aspect word} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

TABLE 1 Examples of SenticNet.

Word	SenticNet(word)
Distrustful	-0.93
Undesirable	-0.35
Likable	0.301
Charming	0.885

$$T^o_{i,j} = \begin{cases} 1 & \text{if } w_i \text{ or } w_j \text{ is a potential opinion word} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

To learn the syntactic information features enhanced by aspect words and opinion words, respectively, we employ two parallel channels. The first channel generates an adjacency matrix that has been augmented by both aspect words and SenticNet affective score, whereas the second channel generates an adjacency matrix that has been enhanced by both opinion words and SenticNet affective score. To effectively integrate the SenticNet affective score with the aspect word weight or opinion word weight, we use the following formula to generate the final enhanced adjacency matrix $A^a_{i,j}$ and $A^o_{i,j}$:

$$W^a_{i,j} = D_{i,j} + S_{i,j} + T^a_{i,j} \tag{7}$$

$$A^a_{i,j} = \frac{1 - e^{-2\times W^a_{i,j}}}{1 + e^{-2\times W^a_{i,j}}} + 0.23841 \tag{8}$$

$$W^o_{i,j} = D_{i,j} + S_{i,j} + T^o_{i,j} \tag{9}$$

$$A^o_{i,j} = \frac{1 - e^{-2\times W^o_{i,j}}}{1 + e^{-2\times W^o_{i,j}}} + 0.23841 \tag{10}$$

When encountering a word that is neither a potential aspect word nor a potential opinion word, and its corresponding SenticNet affective score is 0, the utilization of the bias value of 0.23841 results in an output of 1, with consideration to the precision of five decimal places.

3.5. Feature extraction layer

A two-layer GCN is utilized for contextual feature extraction in each channel. The syntactic dependencies for the potential aspect words or opinion words are captured by feeding the enhanced adjacency matrix $A^a \in \mathbb{R}^{n\times n}$ and the hidden contextual representations $H^c \in \mathbb{R}^{n\times d_l}$ into the GCN module in the left channel. Additionally, the enhanced adjacency matrix $A^o \in \mathbb{R}^{n\times n}$ and the hidden contextual representations $H^c \in \mathbb{R}^{n\times d_l}$ are input into the GCN module of another channel. Inspired by (Zhang et al., 2019), prior to this convolution, we utilize the hidden contextual representations $H^c \in \mathbb{R}^{n\times d_l}$ as input into the multi-target position-aware function $\mathscr{F}^a$ and $\mathscr{F}^o$ to augment the importance of context words close to the potential aspect words or opinion words in two separate channels. Considering that there may be multiple potential aspect terms and opinion terms in one sentence, the function is as follows:

$$q_i^t = \begin{cases} 1 - \frac{\tau+1-i}{n} & 1\leqslant \text{i} < \tau + 1 \\ 0 & \tau + 1 \leqslant \text{i} \leqslant \tau + m \\ 1 - \frac{i-\tau+1-m}{n} & \tau + m < \text{i} \leqslant \text{n} \end{cases} \tag{11}$$

$$\mathscr{F}^a(h_i^l) = \begin{cases} \frac{q_i^1+q_i^2+...+q_i^t}{t} h_i^l & \text{if } w_i \text{ is not a potential aspect word} \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

$$\mathscr{F}^o(h_i^l) = \begin{cases} \frac{q_i^1+q_i^2+...+q_i^t}{t} h_i^l & \text{if } w_i \text{ is not a potential opinion word} \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

where $q_i^t \in \mathbb{R}$ is the position weight to i-th token for the t-th potential aspect term or opinion term in the sentence in two parallel channels, respectively. This function enables the model to effectively avoid noise generated during dependency parsing, resulting in improved performance and more accurate capture of the relevant syntactic dependencies.

The process of GCN is as follows:

$$h_i^l = ReLu(Ag_i^{l-1}W^l + b^l) \tag{14}$$

$$g_i^{l-1} = \mathscr{F}(h_i^{l-1}) \tag{15}$$

where h_i^l denotes the output of the l-th GCN layer. The output of potential aspect term-enhanced GCN layer is $H^a \in \mathbb{R}^{n\times d_l}$, and the output of potential opinion term-enhanced GCN layer is $H^o \in \mathbb{R}^{n\times d_l}$. After that, the final output of Features Extraction Layer H can be computed as follow:

$$H = H^c + H^a + H^o = \{\widetilde{h}_1, \widetilde{h}_2, ..., \widetilde{h}_n\} \tag{16}$$

3.6. Triplet extraction layer

In previous research (Wu et al., 2020a), GTS has been demonstrated to be a highly effective module for extracting triplets from the ASTE task. Therefore, in this study, we have adopted GTS as the decoding algorithm in our proposed model. The output of the Features Extraction Layer is passed through a self-attention layer to extract high-level features. The resulting output is, then, fed into the GTS module. In the GTS module, the relation of two words of the sentence is tagged by set $\{A, O, Pos, Neu, Neg, N\}$. Specifically, the symbols "A" and "O" indicate that the two terms belong to the same triplet, and that they are an aspect term and an opinion term, respectively. The tags "Pos," "Neu," and "Neg" denote the sentiment polarity of the triplet. The symbol "N" represents that there is no association between the two words.An example of the GTS tagging scheme is shown in Figure 4. The following inference strategy is used to predict probability distribution ${p_i{_j}}^t$ of word pair (w_i, w_j) as follows:

$$p_i^{t-1} = maxpooling(p_{i,:}^{t-1}) \tag{17}$$

$$p_j^{t-1} = maxpooling(p_{j,:}^{t-1}) \tag{18}$$

$$q_{ij}^{t-1} = [z_{ij}^{t-1}; p_i^{t-1}; p_j^{t-1}; p_{ij}^{t-1}] \tag{19}$$

$$z_{ij}^t = W_q q_{ij}^{t-1} + b_q \tag{20}$$

$$p_{ij}^t = softmax(W_s z_{ij}^t + b_s) \tag{21}$$

where W_q, W_s, b_q, and b_s are learnable parameters, p_i^{t-1} represents all predicted probability between the word w_i and other words, t denotes the t-th inference, and [.; .] represents the vector concatenation operation. The first three equations are used to observe the probability distribution characteristics of each word pair itself and between word pairs. The initial predicted probability p_{ij}^0 and representation z_{ij}^0 of word pair (w_i, w_j) are set as follows:

$$p_{ij}^0 = softmax(W_s r_{ij}^t + b_s) \tag{22}$$

$$z_{ij}^0 = r_{ij} \tag{23}$$

where $r_{ij} = [\widetilde{h_i}; \widetilde{h_j}]$. Finally, the prediction of the last round is used to extract triplets. The decoding algorithm first predicts aspect terms and opinion terms based on the tags on the main diagonal. It, then, determines whether there are any terms among them that can form a pair. Finally, the most predicted sentiment tag is selected as the sentiment polarity of the pair, and the resulting pair and sentiment polarity are combined to form a triplet.

3.7. Loss function

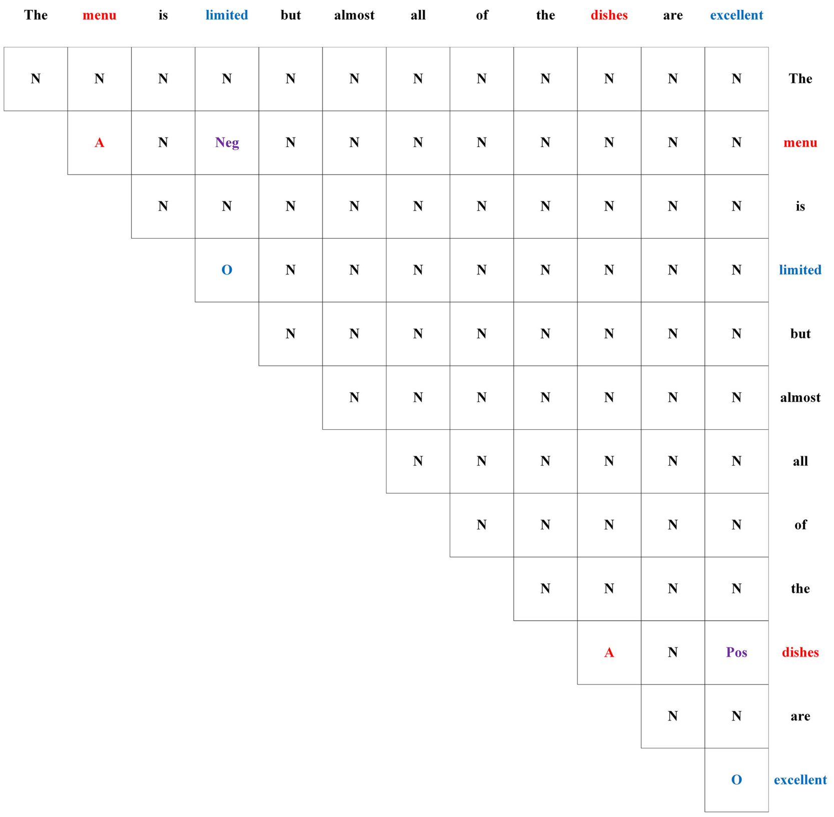
We use the loss function which defined as cross entropy loss between the real label and the predicted label of all word pairs, and the training goal is to minimize it as follows:

$$\mathcal{L} = -\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\in c} I(y_{ij} = k) log(P_{i,j|k}^L) \tag{24}$$

4. Experiments

4.1. Datasets

In this study, we have conducted experiments on three public benchmark datasets from the restaurant domain and a public benchmark dataset from laptop domain named ASTE-Data-V2 mentioned in the study by Xu et al. (2020), all of which have been sourced from the SemEval Challenges and contain 5,989 different comments. Additionally, we have also carried out experiments on the ASTE-Data-V1 datasets mentioned in the study by Wu et al. (2020a) and report the results of these experiments. The details of these datasets are shown in Tables 2, 3.



The	menu	is	limited	but	almost	all	of	the	dishes	are	excellent	
N	N	N	N	N	N	N	N	N	N	N	N	The
	A	N	Neg	N	N	N	N	N	N	N	N	menu
		N	N	N	N	N	N	N	N	N	N	is
			O	N	N	N	N	N	N	N	N	limited
				N	N	N	N	N	N	N	N	but
					N	N	N	N	N	N	N	almost
						N	N	N	N	N	N	all
							N	N	N	N	N	of
								N	N	N	N	the
									A	N	Pos	dishes
										N	N	are
											O	excellent

FIGURE 4
A tagging example with GTS.

TABLE 2 Statistics of the ASTE-Data-V1 datasets.

Datasets	14res			14lap			15res			16res		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
Sentences	1,259	315	493	899	225	332	603	151	325	863	216	328
Triplets	2,356	580	1,008	1,452	383	547	1,038	239	493	1,421	348	525

TABLE 3 Statistics of the ASTE-data-V2 datasets.

Datasets	14res			14lap			15res			16res		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
Sentences	1,266	310	492	906	219	328	605	148	322	857	210	326
Triplets	2,338	577	994	1,460	346	543	1,013	249	485	1,394	339	514

4.2. Evaluation metrics

To ensure the accuracy of the model's performance, Precision (P), Recall (R), and F1 Score (F1) are selected as the evaluation metrics, consistent with prior research in this field:

$$P = \frac{TP}{TP + FP} \tag{25}$$

$$R = \frac{TP}{TP + FN} \tag{26}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{27}$$

where "TP" denotes the number of the positive cases correctly predicted, and "TN" represents the number of negative cases correctly predicted. By contrary, "FP" represents the number of negative cases incorrectly predicted, and "FN" refers to the number of positive cases incorrectly predicted. Notably, the evaluation of extracted triplets is contingent upon the correct prediction of these three components, and any incorrectness in any of these components will render the triplet as incorrect.

4.3. Experiments settings

For the purpose of comparison with previous research, for the Bi-LSTM contextual encoder, following the design of GTS, we use a 300-dimension general-domain embedding from GloVe (Pennington et al., 2014) with 840 billion tokens and a 100-dimension specific-domain embedding from fastText (Bojanowski et al., 2017) to initialize the word embeddings. The hidden state size of the Bi-LSTM is 300, and the dimension is set to 50. The dropout rate of embedding is set to 0.3. For the BERT-based encoder, the bert-base-uncased is used as encoder, and it contains 12 attention mechanism heads, 12 hidden layers, and 768 hidden units. For these two types of encoders, we set Adam optimizer (Kingma and Ba, 2014) to optimize networks with an initial learning rate of 0.001 for the Bi-LSTM contextual encoder and 5e-5 for the BERT-based encoder. The hidden state size of the GCN is set to 300, and the depth of GCN layer is 2. The batch size is set to 32. We conducted 5 independent runs with randomized initialization and reported the experimental results as the average of these five runs.

4.4. Baselines

To evaluate the effectiveness of DGCNAP in the ASTE task, we present other state-of-the-art models in this task for comparison. These models can be categorized into end-to-end models and pipeline models.

Pipeline models

- **CMLA+** (**Peng et al., 2020**) is a two-stage model based on CMLA (Wang et al., 2017). In the first stage, it extracts aspect terms, opinion terms, and sentiment polarities through a multi-layer attention network. In the second stage, it generates possible triplets based on the output of the first stage, then utilizes a binary classifier to filter out invalid triplets.
- **RINANTE+** (**Peng et al., 2020**) is a two-stage model based on RINANTE (Dai and Song, 2019). The only difference between RINANTE+ and CMLA+ is that RINANTE+ extract aspect terms, opinion terms, and triplets through dependency parsing.
- **Li-Unified-R** (**Peng et al., 2020**) is a two-stage framework based on Li-Unified (Li et al., 2019). In the first stage, it uses a customized multi-layer LSTM network to extract targets, opinions, and sentiments. The second stage is similar to CMLA+.
- **Peng + PD** (**Peng et al., 2020**) is a pipeline model. It first predicts all possible triplets, then utilize a MLP classifier to judge the rationality of each triplet.
- **Peng + LOG** (**Wu et al., 2020a**) is a pipeline model. The author add a model proposed in the study by (Fan et al., 2019), after the model proposed in the study by (Peng et al., 2020).
- **IMN-IOG** (**Wu et al., 2020a**) is the combination of the IMM (He et al., 2019) and IOG (Fan et al., 2019) to generate triplets.

End-to-end models

- **OTE-MTL** (**Zhang et al., 2020**) is a model that splits the ASTE task into multiple subtasks, then generate triplets through a bi-affine scorer.
- **JET** (**Xu et al., 2020**) is a unified framework based on the position-aware tagging scheme to generate triplets through an LSTM layer and a CRF layer.
- **GTS** (**Wu et al., 2020a**) is a model that generates triplets by a unified tagging scheme, and the authors design an effective inference strategy to exploit mutual indication between different opinion factors for more accurate extractions.
- **PASTE** (**Mukherjee et al., 2021**) is a tagging-free solution built on an encoder–decoder architecture to produce all triplets.
- **UniASTE** (**Chen et al., 2022**) is a multi-task learning framework which decompose ASTE into three subtasks.
- **GCN-EGTS** (**Hu et al., 2023**) is an end-to-end model which is an enhanced Grid Tagging Scheme (GTS) for ASTE, leveraging syntactic constituency parsing tree and a commonsense knowledge graph based on GCNs.
- **DGEIAN** (**Shi et al., 2022**) is a framework with an interactive attention mechanism. In addition, the authors add different part-of-speech categories in embedding layer.

4.5. Experimental results

The results of our proposed model in the ASTE task are presented in Tables 4, 5. From the results, it is clear that DGCNAP significantly outperforms all other models in terms of F1 score on all datasets. The observations in Table 4 represent that our DGCNAP also performs better than other baseline models on ASTE-Data-V1 datasets. Our method outperforms DGEIAN on the four datasets and acquires 2.36, 1.12, 0.54, and 2.05 improvements in the F1, respectively. Additionally, we observe that the end-to-end model achieves better performance than the pipeline model. For the Bi-LSTM-based encoder, as shown in Table 5, when compared with the best pipeline model, Peng + PD, DGCNAP achieves F1 scores that are more than 10 percentage points higher in three out of the four datasets. On the other hand, in comparison with the model, our proposed model outperforms it by 2.83, 3.7, 1.55, and 3.62 F1 points on the respective datasets. For the BERT-based encoder, DGCNAP also performs well. From the Table 4, it can be observed that the DGCNAP outperforms by 0.06, 4.16, 0.82, and 3.19 F1 points on four datasets when compared with GTS. Our method outperforms the best BERT-based baseline model UniASTE by 1.63, 1.06, 2.14, and 2.36 F1 points, as shown in Table 5. The comparisons presented above demonstrate that our model effectively leverages the affective knowledge information of individual words, leading to improved model's performance in handling sentences with multiple triplets.

4.6. Ablation study

To investigate the effectiveness of the various components in our proposed model, we conducted a series of ablation experiments on the ASTE-data-V2 datasets using the Bi-LSTM encoder. The results of the ablation experiments are presented in Table 6. "w/o SN" refers to the adjacency matrix that is generated only by sentence dependency syntax, without adding SenticNet affective score to the adjacency matrix, and "w/o PA" indicates the model without the multi-target position-aware function in the GCN layer.

TABLE 4 Statistics of the ASTE-Data-V1 datasets.

Encoder	Methods	14res			14lap			15res			16res		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Bi-LSTM	Peng + LOG†	58.89	60.41	59.64	48.62	45.52	47.02	51.70	46.04	48.71	59.25	58.09	58.67
	IMN + IOG†	59.57	63.88	61.65	49.21	46.23	47.68	55.24	52.33	53.75	-	-	-
	GTS-CNN†	70.79	61.70	65.95	55.93	**47.52**	51.38	60.09	**53.57**	56.64	62.63	**66.98**	64.73
	GTS-BiLSTM†	67.28	61.91	64.49	59.42	45.13	51.30	63.26	50.71	56.29	66.07	65.05	65.56
	GCN-EGTS(CNN)	68.74	62.07	65.72	55.94	45.25	49.89	61.54	51.29	55.97	63.73	63.86	63.77
	DGEIAN	71.03	62.63	66.55	60.74	45.56	51.72	**64.87**	52.75	57.11	69.07	65,64	67,30
	DGCNAP	**74.51**	**64.10**	**68.91**	**62.02**	46.09	**52.84**	64.82	51.92	**57.65**	**73.97**	65.29	**69.35**
BERT	GCN-EGTS$_{BERT}$	70.14	68.07	69.20	54.54	52.27	53.64	59.23	**58.15**	58.84	66.89	65.86	66.28
	GTS$_{BERT}$	70.92	**69.49**	70.20	57.52	51.92	54.58	59.29	58.07	58.67	68.58	66.60	67.58
	DGCNAP$_{BERT}$	**71.83**	68.77	**70.26**	**63.91**	**54.34**	**58.74**	**62.03**	57.18	**59.49**	**69.39**	**72.20**	**70.77**

The best results are in bold. The results with "†" are retrieved from the study by Shi et al. (2022), others are retrieved from the original studies.

TABLE 5 Statistics of the ASTE-Data-V2 datasets.

Encoder	Methods	14res			14lap			15res			16res		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Bi-LSTM	CMLA †	39.18	47.13	42.97	30.39	36.92	33.16	34.56	39.84	37.01	41.34	42.10	41.72
	RINANTE +†	31.42	39.38	34.95	21.72	18.66	20.07	29.88	30.06	29.97	25.68	22.30	23.87
	Li-unified-R†	41.04	**67.35**	51.00	40.56	44.28	42.34	44.72	51.39	47.82	37.33	54.51	44.31
	Peng + PD†	43.24	63.66	51.46	37.38	**50.38**	42.87	48.07	**57.51**	52.32	46.96	64.24	54.21
	OTE-MTL†	63.00	55.10	58.70	49.20	40.50	45.10	57.90	42.70	48.90	60.30	53.40	56.50
	JET(M=6)†	61.50	55.13	58.14	53.03	33.89	41.35	64.37	44.33	52.50	70.94	57.00	63.21
	PASTE-AF†	62.40	61.80	62.10	53.70	48.60	51.00	54.80	53.40	54.10	62.20	62.80	62.50
	PASTE-OF†	63.40	61.90	62.60	59.70	48.10	50.00	54.80	52.60	53.70	62.30	63.60	62.90
	UniASTE	70.23	56.82	62.73	55.64	40.91	47.11	63.09	48.37	54.73	66.34	59.26	62.58
	DGEIAN	71.68	61.62	66.26	60.15	43.44	51.14	61.84	50.99	55.89	69.40	60.15	64.37
	DGCNAP	**74.43**	64.49	**69.09**	**64.32**	47.84	**54.84**	**66.73**	50.43	**57.44**	**72.37**	**64.13**	**67.99**
BERT	JET(M = 6)$_{BERT}$	70.56	55.94	62.40	55.39	47.33	51.04	64.45	51.96	57.53	**70.42**	58.37	63.83
	UniASTE$_{BERT}$	72.14	66.30	69.09	**62.24**	51.77	56.51	**64.83**	54.31	59.05	69.06	65.53	67.22
	DGCNAP$_{BERT}$	**72.90**	**68.69**	**70.72**	62.02	**53.79**	**57.57**	62.23	**60.21**	**61.19**	69.75	**69.44**	**69.58**

The best results are in bold. The results with "†" are retrieved from the study by (Shi et al., 2022), others are retrieved from the original studies.

"w/o AE" and "w/o OE" correspond to the models without the aspect words-enhanced GCN channel and the opinion words-enhanced GCN channel, respectively.

Based on the results of the ablation experiments presented in Table 6, we can draw the following conclusion. First, the SenticNet affective score is a crucial component in enhancing the representation of the dependency graph. The utilization of only the adjacency matrix generated from the dependency syntax tree, without incorporating the SenticNet affective score for enhancement, leads to a reduction in the model's ability to predict sentiment polarity. Second, the multi-target position-aware function is another critical module in our proposed model. The removal of this function leads to a significant decrease in the F1 score, the F1 score drops the most to 5.32 on the 14lap dataset, further highlighting the importance of this function in our model. Finally, the ablation experiments reveal that both the aspect terms-enhanced features and the opinion terms-enhanced features are important for model learning. The removal of either of these two channels leads to an average decrease by 0.76 and 1.13 F1 points, emphasizing

TABLE 6 Results of ablation study under the metric of F1.

Model	14res	14lap	15res	16res
DGCNAP	69.09	54.84	57.44	67.99
w/o SN	68.17	52.07	56.98	66.76
w/o PA	64.59	49.52	56.03	64.73
w/o AE	68.59	53.84	57.02	66.89
w/o OE	68.35	53.49	56.57	66.43

TABLE 7 Results of the different usage of SenticNet effective score under the metric of F1.

Model	14res	14lap	15res	16res
DGCNAP	69.09	54.84	57.44	67.99
w/o SN	68.17	52.07	56.98	66.76
DGCNAP-ADD	67.70	51.80	55.96	65.51

their contribution to the overall performance of the DGCNAP model.

4.7. Impact of SenticNet effective score

To investigate the impact of incorporating SenticNet affective score, a series of experiments are conducted on all four ASTE-data-V2 datasets using Bi-LSTM encoder. Specifically, the aim is to explore the impact of using different strategies for incorporating SenticNet effective score. Furthermore, "DGCNAP-ADD" denotes that we generate the final weight of the enhanced graph which is generated by adding the weight of the adjacency matrix to the target score and the absolute value of the SenticNet affective score. The results of the experiments are presented in Table 7, and the corresponding F1 scores are plotted in Figure 5. The experimental results reveal that direct addition of the three values without proper processing during the generation of the final dependency matrix lead to overemphasis of the target words and words with strong emotions. Consequently, the model disregarded the impact of syntactic dependencies and semantic information, leading to undesirable side effects, and resulting in lower performance than the result before adding target weight and SenticNet effective score. Therefore, it is concluded that the incorporation of SenticNet affective score should be carried out with caution as inappropriate usage could have a negative impact on the performance of the model.

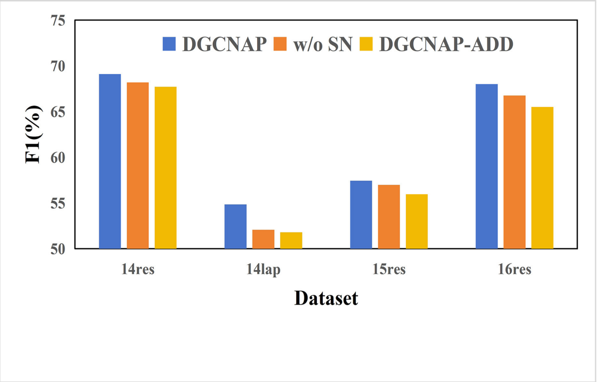


FIGURE 5
F1 scores for different use methods of SenticNet effective score on ASTE-data-V2 datasets.

TABLE 8 Results of the impact of position-aware function study under the metric of F1.

Model	Number of aspect terms			
	1	2	3	4
DGCNAP	66.26	61.70	65.42	43.77
w/o PA	64.44	59.21	63.20	41.81

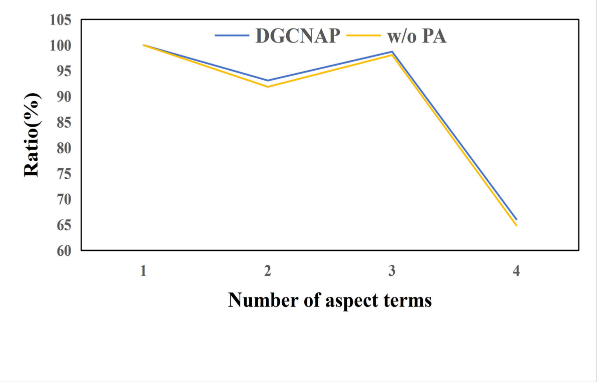


FIGURE 6
The ratio of F1 value of sentences with multiple aspect words to F1 value of sentences with one aspect word.

4.8. Impact of position-aware function

To evaluate the effectiveness of the multi-target position-aware function in sentences with multiple triplets, we conduct experiments on sentences with varying numbers of aspect terms on ASTE-data-V2 datasets using Bi-LSTM encoder. Since the number of sentences with multiple aspect terms in the lap14, res15, and res16 datasets is limited, we conduct experiments on the res14 dataset of ASTE-data-V2 using Bi-LSTM encoder. The experimental results are presented in Table 8, and the ratios of the F1 score value of sentences with multiple aspect terms to the F1 score value of sentences with one aspect term are plotted in Figure 6. The results indicate that the implementation of the multi-target position-aware function has a positive impact on the model's ability to handle sentences with multiple triplets. Specifically, as the number of aspect terms increases, the decline rate of the F1 score value is observed to decrease slower than before implementing the function.

TABLE 9 Results of case study.

Example	Golden truth	GTS	DGCNAP
Once we sailed, the top-notch food and live entertainment sold us on a unforgettable evening.	(Food, top-notch, positive)(Live entertainment, top-notch, positive)	(Food, top-notch, positive)	(food, top-notch, positive) (live entertainment, top-notch, positive)
If you're craving some serious Indian food and desire a cozy ambiance, this is quiet and exquisite choice.	(Ambiance, cozy, positive)(Indian food, serious, positive)	(Ambiance, cozy, positive)(Indian food, serious, positive)(Indian food, craving, positive)	(Ambiance, cozy, positive)(Indian food, serious, positive)
One caveat: Some of the curried casseroles can be a trifle harsh.	(Curried casseroles, neural)	(Curried casseroles, positive)	(Curried casseroles, neural)

4.9. Case study

To show the advantages and disadvantages of DGCNAP, a case study is conducted to compare its performance with that of the GTS model. The results of the study are presented in Table 9. The first sample of the study comprises two triplets, with identical opinion terms. GTS accurately predict only one triplet, while DGCNAP successfully identifies both triplets. The second sample also contains two triplets, but GTS make an erroneous identification of a verb as an opinion term, leading to the prediction of an additional triplet based on the incorrect opinion term. In contrast, DGCNAP accurately recognizes the number of aspect terms and make correct predictions for all triplets. The third sample comprises one triplet. However, due to the fact that GTS does not consider contextual affective knowledge information, it inaccurately determine the sentiment polarity of this triplet. In contrast, DGCNAP accurately predict the sentiment polarity by utilizing the affective knowledge information of each word.

5. Conclusion

This study proposes a novel Dual Graph Convolutional Networks Integrating Affective Knowledge and Position Information (DGCNAP) to the ASTE task, which leverages the contextual features, the affective knowledge information of a single word, and relationship between potential multiple triplets in a same sentence. Specifically, our approach utilize two parallel channels to learn relevant features of potential aspect words and potential opinion words, respectively, by incorporating the SenticNet effective score and the weight of potential aspect words or opinion words when constructing the adjacency matrix. Furthermore, a novel multi-target position-aware function is utilized in the GCN Layer to significantly improve the effectiveness of the model in processing sentences with multiple triplets. The experimental results on four benchmark datasets show the effectiveness of DGCNAP, as it outperforms all other state-of-the-art models significantly in terms of F1 on all datasets. Our analysis on the impact of SenticNet Effective Score and Position-aware Function has demonstrated that these improvements effectively increase the model's ability to identify triplets in sentences. Furthermore, supporting the introduction of affective knowledge can enhance the model's ability to recognize sentiment polarity, while introducing a novel multi-target position-aware function can enhance the interaction between triplets and avoid the impact of noise.

It is noteworthy that one aspect may be associated with multiple opinions and vice versa, and our study has not made improvements to address such situations. For future studies, recognition approaches for handling overlapping triplets will be considered. Additionally, an interactive module will be developed to effectively combine enhancement features of both aspect terms and opinion terms.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

YL: conceptualization. YL and QH: methodology and writing. QH and DZ: funding acquisition and supervision. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

The research was funded by the National Natural Science Foundation of China Research on the Evidence Chain Construction from the Analysis of the Investigation Documents (62166006), supported by the Guizhou Provincial Science and Technology Projects [Guizhou Science Foundation-ZK (2021) General 335], Guizhou Provincial Science and Technology Projects [Guizhou Province under Grant (2020)1Y254], and Guizhou Science and Technology Projects Research on Spatial Optimization and Allocation of Distributed Scientific and Technological Resources [Qiankehe Support(2023) General 093].

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguis.* 5, 135–146. doi: 10.1162/tacl_a_00051

Chen, F., Yang, Z., and Huang, Y. (2022). A multi-task learning framework for end-to-end aspect sentiment triplet extraction. *Neurocomputing* 479, 12–21. doi: 10.1016/j.neucom.2022.01.021

Chen, S., Wang, Y., Liu, J., and Wang, Y. (2021). Bidirectional machine reading comprehension for aspect sentiment triplet extraction. *Proc. Int. AAAI Conf Weblogs.* 35, 12666–12674. doi: 10.1609/aaai.v35i14.17500

Dai, D., Chen, T., Xia, S., Wang, G., and Chen, Z. (2022). Double embedding and bidirectional sentiment dependence detector for aspect sentiment triplet extraction. *Knowledge Based Syst.* 253, 109506. doi: 10.1016/j.knosys.2022.109506

Dai, H., and Song, Y. (2019). "Neural aspect and opinion term extraction with mined rules as weak supervision," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence: Association for Computational Linguistics), 5268–5277.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.* 1, 4171–4186. doi: 10.48550/arXiv.1810.04805

Erik, C., Amir, H., Catherine, H., and Eckl, C. (2009). "Common sense computing: From the society of mind to digital intuition and beyond," in *Biometric ID Management and Multimodal Communication* (Berlin; Heidelberg: Springer), 252–259.

Fan, Z., Wu, Z., Dai, X.-Y., Huang, S., and Chen, J. (2019). Target-oriented opinion words extraction with target-fused neural sequence labeling. *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.* 1, 2509–2518. doi: 10.18653/v1/N19-1259

He, R., Lee, W. S., Ng, H. T., and Dahlmeier, D. (2018). "Exploiting document knowledge for aspect-level sentiment classification," in *Annual Meeting of the Association for Computational Linguistics*, 579–585.

He, R., Lee, W. S., Ng, H. T., and Dahlmeier, D. (2019). "An interactive multi-task learning network for end-to-end aspect-based sentiment analysis," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 504–515. doi: 10.18653/v1/P19-1048

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Hu, Z., Wang, Z., Wang, Y., and Tan, A.-H. (2023). Aspect sentiment triplet extraction incorporating syntactic constituency parsing tree and commonsense knowledge graph. *Cognit. Comput.* 15, 337–347. doi: 10.1007/s12559-022-10078-4

Jebbara, S., and Cimiano, P. (2017). "Improving opinion-target extraction with character-level word embeddings," in *Proceedings of the First Workshop on Subword and Character Level Models in NLP*. Toronto: Association for Computational Linguistics, 159–167.

Jordhy, F., Leylia, K. M., and Akbar, S. A. (2019). "Aspect and opinion terms extraction using double embeddings and attention mechanism for indonesian hotel reviews," in *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)* (Yogyakartha: IEEE), 1–6.

Kingma, D., and Ba, J. (2014). "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations* (San Diego, CA: ICLR), 1051–1060.

Kipf, T., and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv [Preprint]*. arXiv: 1609.02907

Li, X., Bing, L., Li, P., and Lam, W. (2019). A unified model for opinion target extraction and target sentiment prediction. *Proc. Int. AAAI Conf.* 33, 6714–6721. doi: 10.1609/aaai.v33i01.33016714

Li, Y., Lin, Y., Lin, Y., Chang, L., and Zhang, H. (2022). A span-sharing joint extraction framework for harvesting aspect sentiment triplets. *Knowledge Based Syst.* 242, 108366. doi: 10.1016/j.knosys.2022.108366

Liang, B., Su, H., Gui, L., Cambria, E., and Xu, R. (2022). Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. *Knowledge Based Syst.* 235, 107643. doi: 10.1016/j.knosys.2021.107643

Liu, Z., Yang, D., Wang, Y., Lu, M., and Li, R. (2023). Egnn: Graph structure learning based on evolutionary computation helps more in graph neural networks. *Appl. Soft Comput.* 135, 110040. doi: 10.1016/j.asoc.2023.110040

Ma, D., Li, S., Zhang, X., and Wang, H. (2017). "Interactive attention networks for aspect-level sentiment classification," in *Twenty-Sixth International Joint Conference on Artificial Intelligence*, 4068–4074.

Ma, Y., Peng, H., and Cambria, E. (2018). "Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm," in *Proceedings of International AAAI Conference*, Vol. 32 (New Orleans, LA: AAAI Press), 5876–5883. doi: 10.1609/aaai.v32i1.12048

Mukherjee, R., Nayak, T., Butala, Y., Bhattacharya, S., and Goyal, P. (2021). "PASTE: A tagging-free decoding framework using pointer networks for aspect sentiment triplet extraction," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Punta Cana: Association for Computational Linguistics), 9279–9291.

Peng, H., Xu, L., Bing, L., Huang, F., Lu, W., and Si, L. (2020). Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. *Proc. Int. AAAI Conf.* 34, 8600–8607. doi: 10.1609/aaai.v34i05.6383

Pennington, J., Socher, R., and Manning, C. (2014). "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Doha: Association for Computational Linguistics), 1532–1543. doi: 10.3115/v1/D14-1162

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (2014). "Semeval-2014 task 4: aspect based sentiment analysis," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Toronto: Association for Computational Linguistics, 27–35. doi: 10.3115/v1/S14-2004

Shi, L., Han, D., Han, J., Qiao, B., and Wu, G. (2022). Dependency graph enhanced interactive attention network for aspect sentiment triplet extraction. *Neurocomputing* 507, 315–324. doi: 10.1016/j.neucom.2022.07.067

Shi, Y., Li, L., Yang, J., Wang, Y., and Hao, S. (2023). Center-based transfer feature learning with classifier adaptation for surface defect recognition. *Mech. Syst. Signal Process.* 188, 110001. doi: 10.1016/j.ymssp.2022.110001

Tang, D., Qin, B., and Liu, T. (2016). "Aspect level sentiment classification with deep memory network," in *Conference on Empirical Methods in Natural Language Processing* (Austin, TX: Association for Computational Linguistics), 214–224.

Tian, C., Xu, Z., Wang, L., and Liu, Y. (2023). Arc fault detection using artificial intelligence: Challenges and benefits. *Math Biosci Eng.* 20, 12404–12432. doi: 10.3934/mbe.2023552

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*. New York: Curran Associates Inc, 6000–6010.

Wang, W., Pan, S. J., Dahlmeier, D., and Xiao, X. (2017). Coupled multi-layer attentions for co-extraction of aspect and opinion terms. *Proc. Innov. Appl. Artif. Intell. Conf.* 31, 3316–3322. doi: 10.1609/aaai.v31i1.10974

Wang, Y., Liu, Z., Xu, J., and Yan, W. (2023). Heterogeneous network representation learning approach for ethereum identity identification. *IEEE Trans. Comput. Soc. Syst.* 10, 890–899. doi: 10.1109/TCSS.2022.3164719

Wu, H., Zhang, Z., Shi, S., Wu, Q., and Song, H. (2022). Phrase dependency relational graph attention network for aspect-based sentiment analysis. *Knowledge Based Syst.* 236, 107736. doi: 10.1016/j.knosys.2021.107736

Wu, Z., Ying, C., Zhao, F., Fan, Z., Dai, X., and Xia, R. (2020a). "Grid tagging scheme for aspect-oriented fine-grained opinion extraction," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2576–2585. doi: 10.18653/v1/2020.findings-emnlp.234

Wu, Z., Zhao, F., Dai, X.-Y., Huang, S., and Chen, J. (2020b). "Latent opinions transfer network for target-oriented opinion words extraction. *Proc. Innov. Appl. Artif. Intell. Conf.* 34, 9298–9305. doi: 10.1609/aaai.v34i05.6469

Xin, L., Lidong, B., Piji, L., Wai, L., and Zhimou, Y. (2018). "Aspect term extraction with history attention and selective transformation," in *International Joint Conference on Artificial Intelligence*, 4194–4200.

Xu, L., Li, H., Lu, W., and Bing, L. (2020). "Position-aware tagging for aspect sentiment triplet extraction," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2339–2349. doi: 10.18653/v1/2020.emnlp-main.183

Yin, Y., Wei, F., Dong, L., Xu, K., Zhang, M., and Zhou, M. (2016). "Unsupervised word and dependency path embeddings for aspect term extraction," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. California: AAAI Press, 2979–2985.

Zhang, C., Li, Q., and Song, D. (2019). "Aspect-based sentiment classification with aspect-specific graph convolutional networks," in *Conf. Empirical Methods Natural Lang. Process (EMNLP) and the 9th Int. Joint Conf. on Natural Lang. Process (IJCNLP)*, 4560–4570. doi: 10.18653/v1/D19-1464

Zhang, C., Li, Q., Song, D., and Wang, B. (2020). "A multi-task learning framework for opinion triplet extraction," in *Findings of the Association for Computational Linguistics: EMNLP* 2020, 819–828.

Zhang, Y., Ding, Q., Zhu, Z., Liu, P., and Xie, F. (2022). Enhancing aspect and opinion terms semantic relation for aspect sentiment triplet extraction. *J. Intell. Inf. Syst.* 59, 523–542. doi: 10.1007/s10844-022-00710-y

frontiers | Frontiers in Neurorobotics

TYPE Original Research
PUBLISHED 17 August 2023
DOI 10.3389/fnbot.2023.1250706

OPEN ACCESS

EDITED BY
Long Jin,
Lanzhou University, China

REVIEWED BY
Alwin Poulose,
Indian Institute of Science Education and Research, Thiruvananthapuram, India
Mao Yang,
Northeast Electric Power University, China
Yang Li,
Beijing Institute of Petrochemical Technology, China

*CORRESPONDENCE
Shuaishi Liu
✉ liushuaishi@ccut.edu.cn

RECEIVED 30 June 2023
ACCEPTED 31 July 2023
PUBLISHED 17 August 2023

CITATION
Chen Y, Liu S, Zhao D and Ji W (2023) Occlusion facial expression recognition based on feature fusion residual attention network. *Front. Neurorobot.* 17:1250706. doi: 10.3389/fnbot.2023.1250706

Occlusion facial expression recognition based on feature fusion residual attention network

Yuekun Chen, Shuaishi Liu*, Dongxu Zhao and Wenkai Ji

School of Electrical and Electronic Engineering, Changchun University of Technology, Changchun, China

Recognizing occluded facial expressions in the wild poses a significant challenge. However, most previous approaches rely solely on either global or local feature-based methods, leading to the loss of relevant expression features. To address these issues, a feature fusion residual attention network (FFRA-Net) is proposed. FFRA-Net consists of a multi-scale module, a local attention module, and a feature fusion module. The multi-scale module divides the intermediate feature map into several sub-feature maps in an equal manner along the channel dimension. Then, a convolution operation is applied to each of these feature maps to obtain diverse global features. The local attention module divides the intermediate feature map into several sub-feature maps along the spatial dimension. Subsequently, a convolution operation is applied to each of these feature maps, resulting in the extraction of local key features through the attention mechanism. The feature fusion module plays a crucial role in integrating global and local expression features while also establishing residual links between inputs and outputs to compensate for the loss of fine-grained features. Last, two occlusion expression datasets (FM_RAF-DB and SG_RAF-DB) were constructed based on the RAF-DB dataset. Extensive experiments demonstrate that the proposed FFRA-Net achieves excellent results on four datasets: FM_RAF-DB, SG_RAF-DB, RAF-DB, and FERPLUS, with accuracies of 77.87%, 79.50%, 88.66%, and 88.97%, respectively. Thus, the approach presented in this paper demonstrates strong applicability in the context of occluded facial expression recognition (FER).

KEYWORDS

occluded facial expression recognition, feature fusion network, multi-scale module, local attention module, attention mechanism

1. Introduction

Facial expression recognition (FER) has emerged as a critical research direction in the field of artificial intelligence due to the significant role facial expressions play in daily interpersonal communication. FER holds potential applications across diverse fields, including intelligent tutoring systems, service robots, and driver fatigue detection (Poulose et al., 2021a,b). As a result, it has garnered increasing attention in the field of computer vision in recent years.

FER methods can be categorized into two types depending on the scenario: studies conducted in a controlled laboratory environment and studies conducted outside the laboratory in an uncontrolled environment. In controlled environments, the small sample size of the collected data affects the model's feature learning. To overcome this, some researchers propose a new encoder-decoder structure that generates various facial expression images, effectively expanding the sample size (Zhang et al., 2018). Furthermore, Xue et al. (2021) proposed the TransFER model, investigating the relationship between global Transformer-extracted features and local CNN-extracted features. This enhances feature learning and improves model performance. However, these approaches primarily rely on studies conducted on laboratory datasets, such as CK+ (Lucey et al., 2010), MMI (Valstar and Pantic, 2010), and OULU-CASIA (Zhao et al., 2011). Despite achieving high accuracy on these datasets, FER methods exhibit poor performance in uncontrolled environments. To address this, some researchers have tackled class imbalance and label noise issues in datasets by utilizing techniques like data augmentation and auxiliary datasets (Wang et al., 2018). Network interpretability studies demonstrate that models can prioritize relevant facial expression features, resulting in more accurate emotion detection (Kim et al., 2021). Additionally, the noisy labeling problem in real-world datasets can be mitigated by introducing a probabilistic transformation layer (Zeng et al., 2018). The above methods are investigated on expression datasets in uncontrolled environments. However, FER still faces challenges when the face is partially occluded by objects like sunglasses, scarves, masks, or other random items that frequently occur in real images or videos.

Addressing the facial occlusion problem is crucial for improving the performance of FER models in real-world environments. As shown in Figure 1, the occlusion problem leads to a large spatial change in the appearance of the face. To tackle this issue, certain researchers have suggested utilizing deep CNN networks for solving the occlusion problem. Specifically, two CNN networks are trained from a global perspective using occluded and non-occluded face images. The non-occluded face images are utilized as privileged information for fine-tuning the occluded expression recognition network. This approach (Pan et al., 2019) significantly reduces occlusion interference and enhances network performance. However, the drawback of this FER algorithm is its focus solely on global features, neglecting the crucial local detail features that play a vital role in expression discrimination. Therefore, regarding the occlusion FER problem, certain researchers suggested a method based on local keypoint localization (Wang K. et al., 2019), effectively capturing crucial local facial features. However, choosing the appropriate local regions remains a key issue. To address this, researchers employed three local region generation schemes: fixed position selection, random selection, and labeled keypoint selection. This approach significantly enhances the performance of the occlusion FER model. An alternative method for keypoint selection involves choosing 24 facial keypoints to define 24 key local regions. Subsequently, an attention network is employed to extract features from each region, allowing better focus on important local features. This approach (Li et al., 2018) offers a viable solution to the occlusion FER problem. Nonetheless, the localization-based approach has a drawback of neglecting global information, which limits its overall ability in expression discrimination. Consequently, the effective combination of global and local features is paramount in addressing the occlusion FER problem.

To solve the above issues, a feature fusion residual attention network aiming to enhance feature robustness is proposed. In convolutional neural networks (CNNs), deep convolutions exhibit a broader receptive domain and encompass richer semantic features, whereas shallow convolutions have a narrower receptive domain and capture rich profile features. However, deep convolutions are susceptible to occlusion (Proverbio and Cerri, 2022). To address this, this paper employ multi-scale modules to extract features from diverse receptive domains, thereby enhancing the diversity and robustness of global features. Additionally, this paper design local attention modules to extract local features, mitigating occlusion interference. To learn both global multi-scale and local features, this paper employed a two-branch network. The first branch utilized the multi-scale module, while the second branch divided the extracted feature maps into multiple non-overlapping local feature maps, which were then processed using the attention mechanism. Finally, the processed features were fused. The main contributions of this paper can be summarized as follows:

1. Feature fusion residual attention network (FFRA-Net), a simple and effective FER network, is proposed to address the challenge of facial occlusion by enhancing the diversity of expression features through feature fusion.
2. The multi-scale module extracts features at different scales from the feature map, thereby reducing the sensitivity of deep convolutions to occlusion. Additionally, the local attention module focuses on local salient features and mitigates occlusion interference.

The remainder of this paper is structured as follows. Section 2 provides a review of relevant literature. Subsequently, the proposed approach is presented in Section 3. Section 4 presents the experimental results for both obscured and non-obscured expression datasets. Additionally, visualizations are provided to further validate the proposed method. Section 5 summarizes the findings.

2. Related work

2.1. Deep convolutional FER

In recent years, researchers have made significant progress in FER by proposing numerous methods based on deep CNNs. However, deep learning-based FER often disregards domain-specific knowledge related to facial expressions. To tackle this issue, Chen et al. (2019) introduced a framework for FER that leverages prior knowledge by utilizing the distinctions between neutral expressions and other expressions to train the network. Moreover, head pose variation poses a common challenge in expression recognition. To tackle this issue, Marrero-Fernández et al. (2019) propose an end-to-end architecture with an attention mechanism that rectifies facial images to improve expression classification. Due to the subtle variations in expressions, the issue of inter-class similarity in expression datasets becomes crucial. To address this, Wen et al. (2021) proposed attention distraction networks. The aforementioned methods primarily concentrate on datasets obtained in controlled environments, where facial images are predominantly frontal. Consequently, the model's performance suffers when it comes to recognizing facial expressions in uncontrolled environments.

To differentiate between uncertain and blurred expression images in uncontrolled environments, Pu et al. (2020) proposed an expression recognition framework based on facial action units. The framework incorporates an attention mechanism that

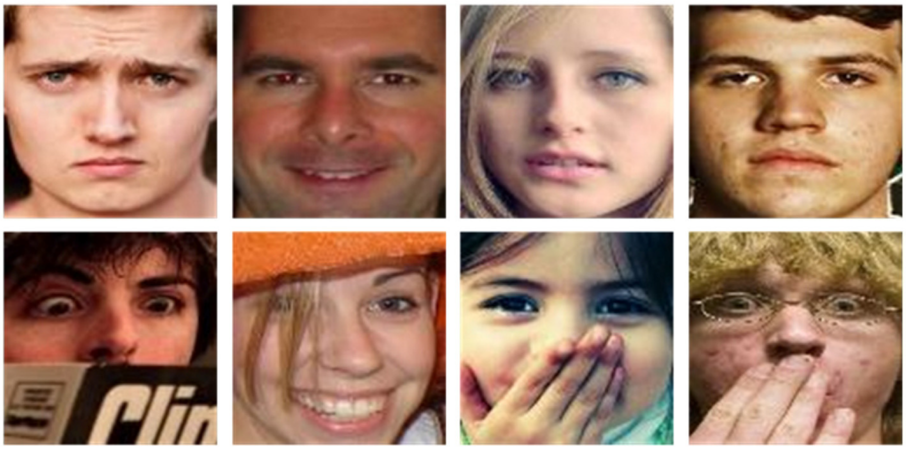


FIGURE 1
Some examples of images from the RAF-DB dataset, where the first row comprises non-occluded expression images and the second row comprises occluded expression images.

dynamically focuses on significant facial actions. To quantify these uncertainties, Zhang et al. (2021) proposed a relative uncertainty method that assigns weights based on uncertainties, fuses facial features, and introduces a new uncertainty loss. She et al. (2021) introduced a multi-branch learning network to address the label ambiguity problem in FER. The method enhances the ability to explore and capture the underlying distribution in the label space. Furthermore, the expression dataset faces challenges posed by pose variation and identity bias. To tackle these challenges, Wang C. et al. (2019) proposed an adversarial feature learning method. The gesture discriminator and identity discriminator classify gestures and identities based on the extracted feature representations, respectively. Similarly, Chen and Joo (2021) presented a FER framework based on facial action units. The framework integrates a triple loss into the objective function, leading to improved expression classification accuracy. Despite the impressive performance of the aforementioned methods on uncontrolled environment data, the task of masking FER remains challenging.

2.2. Occluded FER

Considering the limited availability of large-scale occluded expression datasets, Xia and Wang (2020) proposed a stepwise learning strategy for occluded FER models. The distribution density in the feature space is first used to measure the complexity of the non-occluded data, thus guiding the distribution of the occluded expression features to converge to the distribution of the non-occluded expression features. In a similar vein, Pan et al. (2019) presented a novel method for occluded FER that leverages non-obscured face image information. This approach aims to align the distribution of learned occluded face image features with the distribution of non-occluded face image features. Nonetheless, the aforementioned methods rely on global features. In occlusion expression recognition, global features are susceptible to the influence of occlusion, leading to reduced accuracy in expression recognition. To overcome this challenge, Wang K. et al. (2019) introduced a network based on local region attention. Additionally, they proposed a region bias loss to assign weights to local region attention. Xue et al. (2022) proposed a dedicated attention mechanism for FER networks. The proposed model selectively focuses on the most relevant expression features while disregarding irrelevant features, thereby avoiding undue emphasis on occlusion or other noisy regions. The aforementioned approach based on local features effectively addresses the occlusion problem. However, it overlooks global information and possesses limited discriminative ability for expression as a whole.

Hence, it is crucial to consider both global and local features for effective occluded expression recognition. Ding et al. (2020) introduced an adaptive depth network for recognizing occluded facial expressions. Initially, global features are extracted using the ResNet-50 backbone network. Subsequently, the network is partitioned into two branches. Each branch is further divided into multiple sub-regions, with each sub-region independently predicting expressions. Finally, strategy fusion is conducted to obtain the final classification results. Zhao et al. (2021) presented an expression recognition network capable of learning global and local features. This network effectively mitigates the deep network's sensitivity to occlusion and autonomously attends to local key information. Finally, the same policy fusion is employed to derive the results. Nevertheless, the policy fusion approach is prone to overfitting as the network deepens and shows poor performance when trained on certain realistic occlusion data.

3. Proposed method

FFRA-Net is a feature fusion network designed to address the recognition of obscured facial expressions. The method comprises a multi-scale module, a local attention module, a feature fusion module, and a residual link. The backbone network chosen for this purpose is ResNet-18 (He et al., 2015). Figure 2 illustrates the structure of FFRA-Net. Initially, the feature preextractor captures the intermediate facial expression features, which are obtained from the first three convolutional stages of ResNet-18. Then, a two-branch network is used to process the acquired intermediate feature maps into the multiscale module and the local attention module, respectively, allowing the model to obtain both global and local expression features. Subsequently, the model enters the feature fusion phase, where a weighted fusion approach is applied to assign specific weights to the feature mappings from the two branches. These weighted features are then directly summed. Meanwhile, it is then added with the original intermediate feature map to form a residual connection, and finally a global and local attention feature map is obtained. Finally, this feature map proceeds to the last convolutional stage of ResNet-18, followed by fully connected layers for deriving the classification results.

3.1. Multi-scale module

Multi-scale modules are widely used in computer vision for processing visual information across different scales (Gao et al., 2019; Ma and Zhang, 2023). It is widely used in many tasks, including target detection and image segmentation. Typically, the multi-scale module divides the feature map into multiple subregions of different scales in the spatial dimension, processing each subregion individually. However, this approach is primarily applicable to visual tasks like target detection and image segmentation. Occluded expression recognition is influenced by occlusions, leading to the absence of certain semantic information. To compensate for this deficiency, there is a need for more comprehensive and diverse global features. To tackle this issue, a novel multi-scale image classification module is proposed (Figure 3A). The feature map is divided into multiple sub-feature maps along the channel dimension, enabling the extraction of a broader range of global expression information.

The objective of this method is to learn multi-scale features within the feature map while ensuring that the feature subsets encompass a wider range of scale information. Specifically, the feature mapping X is obtained through feature pre-extraction. Next, the module partition X into n feature map subsets along the channel axis, denoted as X_i, with $i \in \{1, 2, \ldots, n\}$ representing the index. Each feature subset X_i has the same spatial size as the feature map X but contains only $1/n$ channels. Subsequently, a 3×3 convolution is applied to each X_i, yielding the output denoted as P_i^{ms}, while Y_i^{ms} represents the output after fusion of each sub-feature. Therefore, the expression for each output Y_i^{ms} can be defined as follows:

$$Y_i^{ms} = \begin{cases} P_i^{ms}(X_i) & i = 1 \\ P_i^{ms}\left(X_i + Y_{i-1}^{ms}\right) & 1 < i \leqslant n \end{cases} \tag{1}$$

Equation 1 demonstrates that each output Y_i^{ms} encompasses a distinct number and scale of subset features. In order to obtain a more diverse collection of global features, the module concatenate all the Y_i^{ms} outputs along the channel dimension. However, increasing the value of n results in features containing more scale information, which in turn increases model complexity and computational overhead. Taking these factors into consideration, n is set to 4 in this module to optimize the performance of the model.

The multi-scale convolution captures comprehensive and detailed global information in the feature map, thereby reducing the sensitivity of deep convolution to occlusion. Compared to the traditional ResNet-18 network, this network selectively attends to the facial regions related to expression while disregarding occluded regions, thus effectively addressing the issue of facial occlusion.

3.2. Local attention module

The local attention module, commonly used in computer vision, utilizes the attention mechanism to capture essential information from images. The attention mechanism, similar to human vision, assigns weights to channels or spatial domains through automatic learning. This enables the neural network to focus on important regions and disregard others. In occlusion-based FER, a portion of the facial image is obscured by an occluder, leading to a loss of discriminative ability in the occluded region's features. Based on this feature, a novel local attention module (Figure 3B) is proposed. This module significantly enhances the model's perceptual capability.

Local features play a crucial role in occlusion FER. However, previous methods often employ face tagging or random cropping to divide faces into multiple local regions in order to extract effective local features. but these methods may result in redundancy of features and increase in computational overhead. To solve this issue, the intermediate feature maps are divided into non-overlapping local feature maps, aiming to enable each local feature map to autonomously focus on local key features using attention mechanism. Therefore, after 3×3 convolution of the feature maps obtained by feature pre-extraction, the module divide the extracted feature map S into several local feature maps S_i along the spatial axis, where $i \in \{1, 2, \ldots, m\}$. Each S_i undergoes a 3×3 convolution, resulting in a feature map denoted as $F \in \mathbb{R}^{H \times W \times C}$. Shuffle Attention (SA) mechanism was subsequently used as the attention network (Zhang and Yang, 2021). The SA module divides the input feature map into G sub-feature maps evenly across the channel dimension, where G is set to 8. Subsequently, each sub-feature map is evenly divided into two feature maps along the channel dimension. Then, the SA module calculates the channel and spatial attention weights for each of the two feature maps successively, focusing on the channel and spatial dimensions, respectively. Subsequently, the attention weights are multiplied with the original feature maps to generate attention maps in both dimensions. As shown in Equations 2 and 3, these two attention maps are then combined, and the same process is repeated for the remaining sub-feature maps. The interaction between each sub-feature graph is achieved through the channel shuffle operation. Channel shuffle involves randomly rearranging the original channel order of the feature map before their combination. Finally, an attention graph with the same shape as the input feature graph is generated. In our network, each $F_i \in \mathbb{R}^{H \times W \times C/G}$ (where $i \in \{1, 2, \ldots, G\}$) is further divided into $F_{ij} \in \mathbb{R}^{H \times W \times C/2G}$ (where $j \in \{1, 2\}$), and the attention network takes F_{ij} as input. It calculates a one-dimensional channel attention weight map $M_c \in \mathbb{R}^{1 \times 1 \times C}$ and a two-dimensional spatial attention weight map $M_s \in \mathbb{R}^{H \times W \times 1}$ for element-level multiplication denoted by $\otimes$, and outputs the result

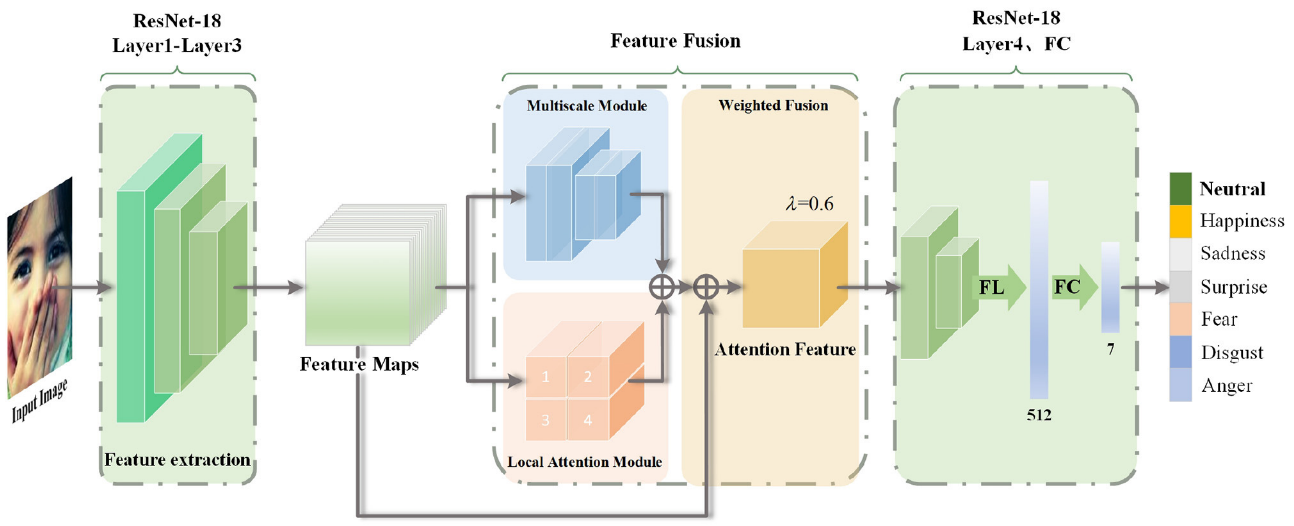


FIGURE 2
The model structure employed in this paper. The method includes the backbone network ResNet-18, along with a multi-scale module, a local attention module, and a weighted fusion module. Here, Input image from the RAF-DB dataset, λ = 0.6 represents the weight assigned to local features, FL indicates the tiling operation, and FC refers to the fully connected network.

as F_r after stitching the sub-attention maps (F_r). Therefore, the attention network can be expressed as follows:

$$F_{ri} = \left[\left(M_s \left(F_{ij} \right) \otimes F_{ij} \right), \left(M_c \left(F_{ij} \right) \otimes F_{ij} \right) \right] \tag{2}$$

$$F_r = [\mathrm{F}_{\mathrm{r}1}, \cdots, \mathrm{F}_{\mathrm{rG}}] \tag{3}$$

Let the output of the 3 × 3 convolutional and attentional network be denoted as P_i^{la}, and the output after feature fusion as Y_i^{la}. Thus, each output can be expressed as follows:

$$Y_i^{la} = \begin{cases} P_i^{la}\left(S_i\right) & i = 1 \\ P_i^{la}\left(S_i + Y_{i-1}^{la}\right) & 1 < i \leqslant n \end{cases} \tag{4}$$

Based on Equation 4, each output comprises varying numbers and sizes of local features. To obtain a wider range of diverse local features, the module concatenate all the outputs along the spatial dimensions. In this study, m is set to 4, which aligns better with the characteristics of masked expression images and guarantees improved model performance.

3.3. Feature fusion module

In computer vision, a feature fusion module is employed to integrate information from diverse feature types, enhancing the performance of vision tasks. To maintain a balance between the significance of multi-scale and local attention features, weights are incorporated into the feature fusion module. Figure 3C illustrates the integration of global and local information within this module, resulting in improved model performance. Furthermore, to enhance the network's expressive capacity, the module establish residual connections between the input and output features. This enables the network to more effectively capture image details and contextual information. Here, the original input features are denoted as X, the outputs of the multi-scale and local attention modules as Y_i^{ms} and Y_i^{la}, respectively, and the output of the final feature fusion module as X. Therefore, it can be expressed as:

$$Y = \lambda Y_i^{\mathrm{la}} + (1 - \lambda) Y_i^{\mathrm{ms}} + X \tag{5}$$

In Equation 5, λ represents a hyperparameter that controls the relative significance of the multi-scale and local attention modules. It is demonstrating experimentally that the model achieves the best performance when λ is set to 0.6.

4. Experiment

This section describes the data set used and the data processing procedures. And the details of the experimental setup are presented. Then, the experimental results are presented, including the results of the ablation experiments, the determination of the feature fusion weights, the visualization of the CAM, and the results of the partial confusion matrix. Last, the method of this paper is compared with other methods, and the experimental results are comprehensively analyzed.

4.1. Datasets

RAF-DB (Li and Deng, 2019): RAF-DB, a real-world expression dataset, comprises 29,672 facial expression images. These images were independently annotated by approximately 40 annotators. The experiments in this paper utilized a single tag provided by RAF-DB. The dataset consists of 15,339 expression images, encompassing six basic expressions (happy, surprised, sad, angry, disgusted, and fearful), as well as neutral expressions. Out of

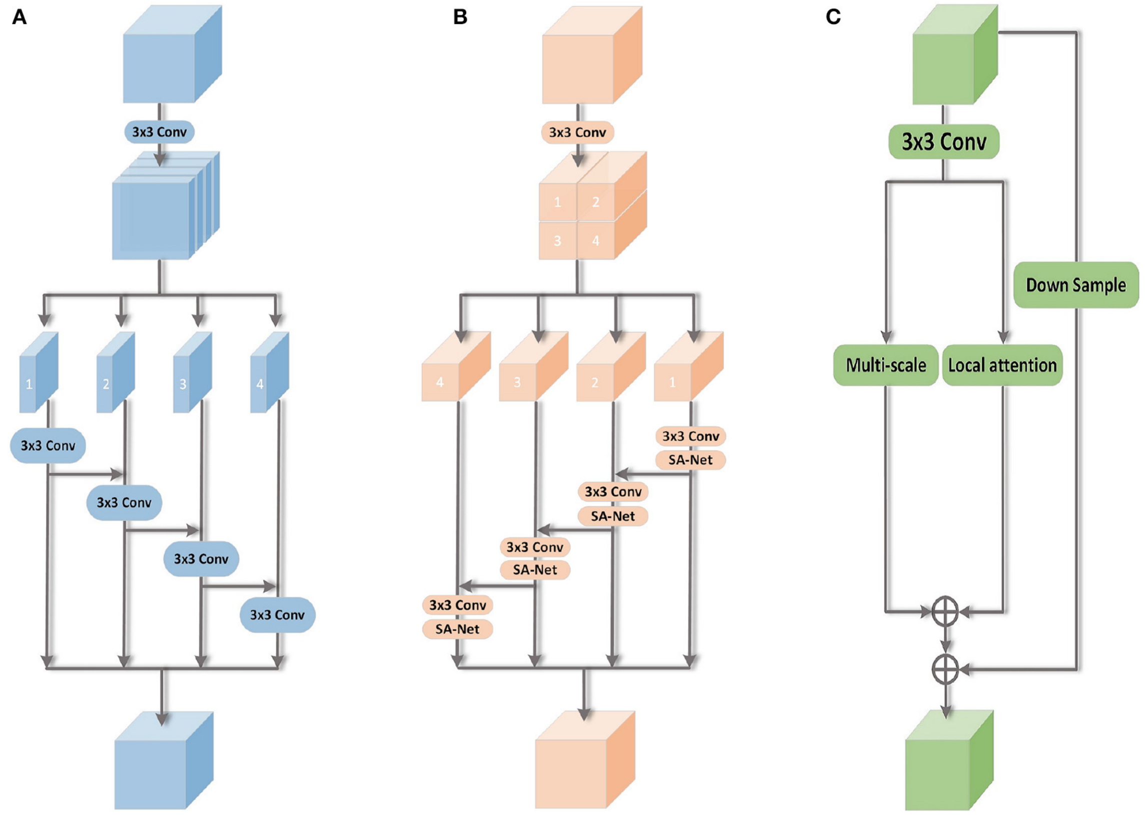


FIGURE 3
FFRA-Net uses three types of modules. Multi-scale module, local attention module, and feature fusion module. **(A)** Multi-scale. **(B)** Local attention. **(C)** Feature fusion.

these, 12,271 images were allocated for training, while 3,068 were allocated for testing.

FERPLUS (Barsoum et al., 2016): FERPLUS is an extension of FER2013, a large-scale dataset collected using the Google Image Search API. The dataset comprises 28,709 training images, 3,589 validation images, and 3,589 test images. It was re-labeled by 10 annotation workers to include six basic expressions (happy, surprised, sad, angry, disgusted, and fearful), as well as neutral and contemptuous expressions.

FM_RAF-DB and SG_RAF-DB: To evaluate the performance of our proposed FER model under realistic occlusion conditions, two occlusion representation datasets were created based on RAF-DB: FM_RAF-DB and SG_RAF-DB. Using face detection (Deng et al., 2020), these datasets simulate both cases of faces wearing masks and sunglasses. The masked face method used, specifically, marks the key points of the face and selects the key points around the eyes and mouth. The method then uses a bionic matrix and a bionic transformation calculation to place the mask image and the sunglasses image in their respective positions (refer to Figure 4). These two datasets better simulate the facial occlusion in real scenes, allowing a more accurate evaluation of the performance of our proposed FER model.

4.2. Implementation details

For all datasets, official face-aligned samples are used. The input images of RAF-DB and FERPLUS datasets were cropped to a size of pixels, respectively. In this study, the ResNset-18 network was chosen as the backbone network and the experimental code was implemented using the PyTorch framework. The training was conducted on an NVIDIA RTX-3090 GPU. In this study, a pre-trained ResNet-18 model obtained by training on the MS-Celeb-1M dataset was utilized. The optimizer used for training is the Adam optimizer with a batch size of 128 and an initial learning rate of 0.0001. To achieve the best results, the model in this paper was trained on all datasets for 200 epochs.

4.3. Ablation studies

In order to assess the effectiveness of FFRA-Net, this section performed ablation experiments on the FM_RAF-DB and SG_RAF-DB datasets. The experimental results encompass the selection of feature fusion strategy, the value of the weight hyperparameter, the impacts of the multi-scale

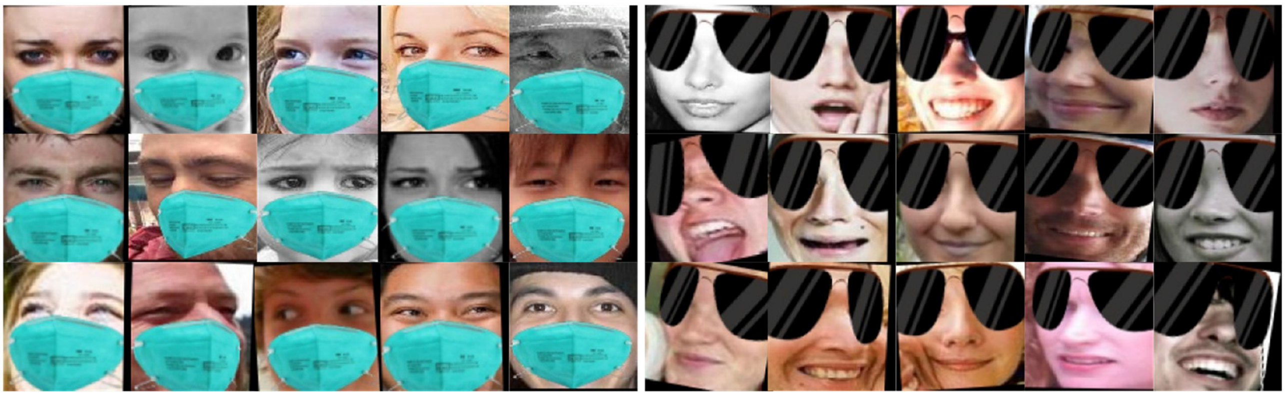


FIGURE 4
Some image examples of FM_RAF-DB and SG_RAF-DB dataset.

TABLE 1 Evaluating various fusion strategies on the SG_RAF-DB dataset.

Fusion strategies	Acc.(%)
Concate feature fusion	76.69
Add feature fusion	77.87
Weighted feature fusion	79.50

module and local attention module on the model, as well as CAM visualization.

4.3.1. Selection of the feature fusion strategy

In this subsection, different fusion strategies are experimented on the SG_RAF-DB dataset. Table 1 presents the comparison results of three feature fusion strategies: splicing fusion, summing fusion, and weighted fusion. Splicing fusion involves concatenating two feature maps along the channel dimensions and subsequently fusing the information from all channels through convolution. Additive fusion directly adds the feature maps obtained from two branches to create a combined feature map. Weighted fusion assigns specific weights to the feature maps of different branches based on additive fusion and then adds them together. In this study, the weight for the local attention module is empirically set to 0.6, as verified in subsequent subsections. The results demonstrate that weighted fusion is a more suitable fusion method.

4.3.2. The value of the weight hyperparameter λ

To balance the importance of multi-scale modules and local attention modules, λ is used as a hyperparameter. The local attention weight is set to λ, and the weight of the multiscale module is set to 1 − λ. This experiment investigate different values of λ ranging from 0.1 to 0.9 to examine its effect on FFRA-Net, and the results are presented in Figure 5. When λ is set to 0.6, the weight of the local attention branch is slightly higher than that of the multi-scale branch, leading to the model achieving the best performance.

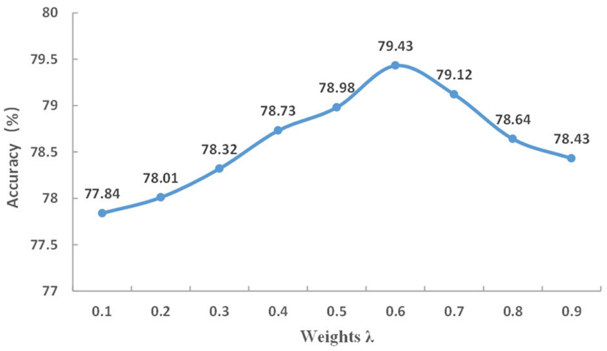


FIGURE 5
Evaluation of different λ values on the SG_RAF-DB dataset.

TABLE 2 Evaluation of multi-scale and local attention modules in networks on the FM_RAF-DB and SG_RAF-DB datasets.

Multi-scale	Local attention	FM_RAF-DB	SG_RAF-DB
-	-	75.98%	77.44%
✓	-	76.86%	78.62%
-	✓	77.74%	79.37%
✓	✓	77.87%	79.50%

4.3.3. Effects of multi-scale modules and local attention modules

An ablation analysis was conducted to verify the effectiveness of the multi-scale module and the local attention module in FFRA-Net. The results in Table 2 demonstrate that using either the multi-scale module or the local attention module alone yields higher accuracy compared to the baseline accuracy. Moreover, the local attention module exhibits greater usefulness than the multi-scale module. Ultimately, the model achieved the best performance by employing both modules and integrating their features.

To provide a clearer understanding of the effect of the feature fusion module, the study conducted CAM visualization (Zhou et al.,

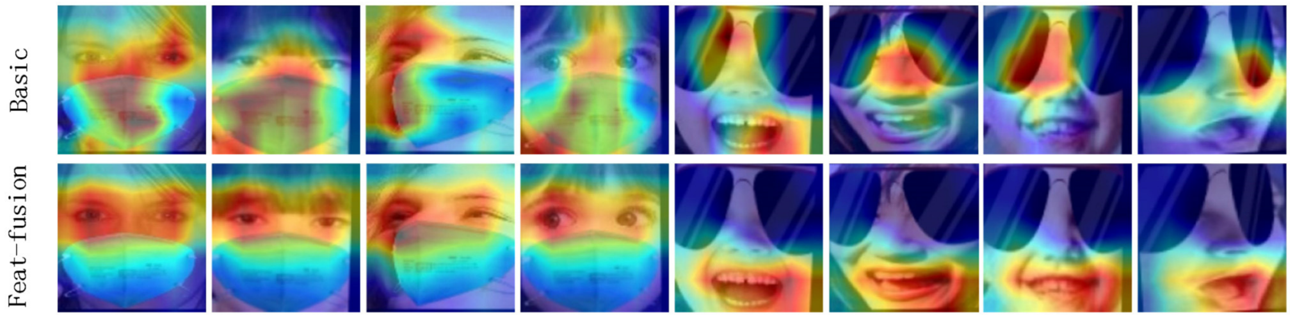


FIGURE 6
Feature fusion and CAM visualization of ResNet-18. Images are from the test set of FM_RAF-DB and SG_RAF-DB datasets.

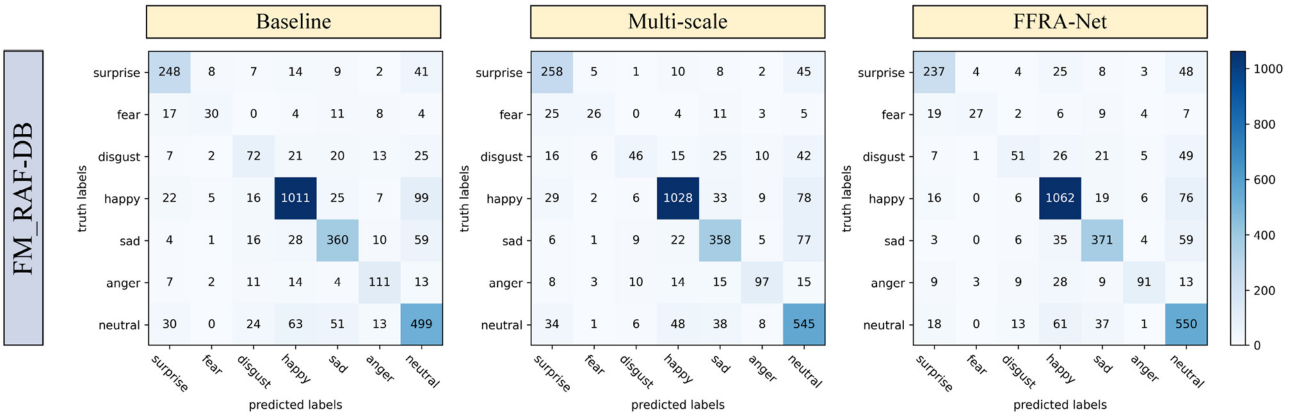


FIGURE 7
Confusion matrix results for baseline, multi-scale modules and FFRA-Net on the FM_RAF-DB test set.

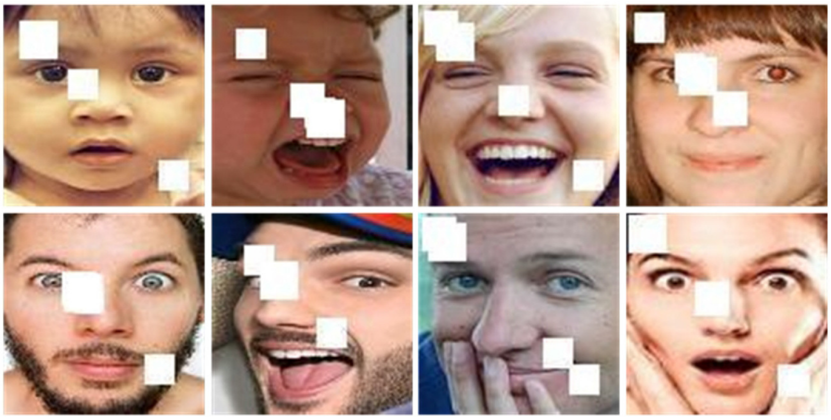


FIGURE 8
The images were captured from the test set of the RAF-DB dataset, augmented with random occlusion.

2015) to validate its performance. Figure 6 displays the visualization results of the baseline and feature fusion modules in the first and second rows, respectively. In comparison to the traditional ResNet-18, the CAM results obtained with feature fusion direct the network's attention toward locally significant regions. For the first four images where faces are covered by masks, even though the mouth is the primary region of the mask, the model predominantly focuses on the eye region. Similarly, for the last four images where faces are covered by sunglasses, despite the eye being the main region of the mask, the model primarily attends to the mouth region, which aligns with human perception. The results indicate that methods in this paper effectively addresses the occlusion problem.

4.4. Confusion matrix analysis

Confusion Matrix is a valuable tool for evaluating the performance of a classification model. It displays the relationship between the classification model's predictions for different categories and their corresponding true labels, with the table numbers representing the number of predicted samples. The subsection analyze the Confusion Matrix of the baseline, multi-scale module, and FFRA method applied to the test set of the FM_RAF-DB dataset. Figure 7 displays the Confusion Matrix. FFRA Method significantly improves the recognition accuracy of the neutral expression category. Neutral expressions, being states without obvious emotional signals, may lack distinct facial expression features compared to other expression categories. However, FFRA Method can effectively focus on more accurate and relevant features when recognizing neutral expressions, thereby enabling the model to achieve higher recognition accuracy.

4.5. Assessment of the model's performance in real-world scenarios

To further validate the performance of the FFRA model in real-world environments, the test set of the RAF-DB dataset was added with random occlusion, as depicted in Figure 8. The model achieves an accuracy of 86.43% on this dataset, surpassing the performance of other FER methods listed in Table 3. This demonstrates the outstanding performance of the model in real-world scenarios.

4.6. Comparison with previous results

In this section, FFRA-Net is compared with other state-of-the-art methods using the FM_RAF-DB and SG_RAF-DB datasets. Specifically, VGG-16 (Simonyan and Zisserman, 2014), ResNet-50 (He et al., 2015), and MobileNetv2 (Sandler et al., 2018) are models with larger parameter counts, deeper networks, and lighter weights, respectively, while SCN (Wang et al., 2020) and MA-Net (Zhao et al., 2021) are specifically designed for FER in the wild. The experimental results in Table 4 demonstrate that FFRA-Net outperforms the other FER models in terms of accuracy, showcasing excellent performance.

FFRA method achieves an accuracy of 77.87% on the FM_RAF-DB dataset and 79.50% on the SG_RAF-DB dataset. These results surpass several existing mainstream methods and occluded FER methods. The proposed FFRA-Net in this paper exhibits outstanding performance in recognizing obscured expression images.

The accuracy results of FFRA-Net and other FER models on the RAF-DB and FERPLUS datasets are shown in Table 5. The FFRA method achieves an accuracy of 88.66% on the RAF-DB dataset and 88.97% on the FERPLUS dataset. These results outperform several existing FER methods in the wild. The results demonstrate that the proposed method in this paper exhibits strong generalization ability.

FFRA method achieves an accuracy of 88.66% on the RAF-DB dataset and 88.97% on the FERPLUS dataset. These results surpass several existing expression recognition methods. The results show that the method proposed in this paper has a strong generalization ability.

TABLE 3 Comparison of performance with previous FER methods on the test set of the RAF-DB dataset after incorporating random occlusion.

Method	Acc.(%)
Baseline	81.62
SCN (Wang et al., 2020)	85.78
MA-Net (Zhao et al., 2021)	86.23
FFRA-Net (Ours)	**86.43**

The bold values are outcomes from model runs described in this paper.

TABLE 4 Performance comparison (%) with previous methods on FM_RAF-DB and SG_RAF-DB.

Methods	FM_RAF-DB	SG_RAF-DB
VGG-16 (Simonyan and Zisserman, 2014)	73.86	75.81
ResNet-50 (He et al., 2015)	74.32	75.88
MobileNetv2 (Sandler et al., 2018)	73.14	75.46
SCN (Wang et al., 2020)	76.43	77.64
MA-Net (Zhao et al., 2021)	77.64	78.78
FFRA-Net(Ours)	**77.87**	**79.50**

The bold values are outcomes from model runs described in this paper.

TABLE 5 Performance comparison (%) with previous methods on RAF-DB and FERPLUS.

Methods	RAF-DB	FERPLUS
gACNN (Li et al., 2019)	85.07	-
RAN (Wang K. et al., 2019)	86.90	88.55
SCN (Wang et al., 2020)	87.03	88.01
DACL (Farzaneh and Qi, 2021)	87.78	-
KTN (Li et al., 2021)	88.07	-
MA-Net (Zhao et al., 2021)	88.40	-
RUL (Zhang et al., 2021)	-	88.75
DMUE (She et al., 2021)	-	88.64
SeNet50 (Albanie et al., 2018)	-	88.80
FFRA-Net(Ours)	**88.66**	**88.97**

The bold values are outcomes from model runs described in this paper.

5. Conclusion

To solve the problem of occluded FER, a new feature fusion architecture, called FFRA-Net, is proposed, which can learn a rich diversity of global and local features. First, a multi-scale module is proposed to provide diverse global features. Second, an attention-based mechanism local attention module is proposed, which assigns higher weights to important facial regions and smaller weights to irrelevant facial regions. Finally, a feature fusion module is proposed, which uses a weighted approach to fuse global and local features. Extensive experiments on four FER datasets show that this method outperforms the existing FER methods. However, the model requires further optimization in terms of parameter reduction to alleviate computational overhead. A primary area of future research is the investigation of lightweight techniques for occluded FER.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

Data curation: YC. Conceptualization: YC, SL, and DZ. Methodology, software, writing—original draft, and validation: YC and DZ. Formal analysis: YC, SL, and WJ. Supervision: SL. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the Project of National Natural Science Foundation of China under Grant No. 62106023.

Acknowledgments

The authors would like to thank the reviewers and the Technical Editor for their valuable comments and suggestions on revising this paper.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Albanie, S., Nagrani, A., Vedaldi, A., and Zisserman, A. (2018). "Emotion recognition in speech using cross-modal transfer in the wild," in *Proceedings of the 26th ACM international conference on Multimedia, pages* (New York, NY), 292–301. doi: 10.1145/3240508.3240578

Barsoum, E., Zhang, C., Canton-Ferrer, C., and Zhang, Z. (2016). "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (New York, NY), 279–283. doi: 10.1145/2993148.2993165

Chen, Y., and Joo, J. (2021). "Understanding and mitigating annotation bias in facial expression recognition," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (Montreal, QC: IEEE), 14960–14971. doi: 10.1109/ICCV48922.2021.01471

Chen, Y., Wang, J., Chen, S., Shi, Z., and Cai, J. (2019). "Facial motion prior networks for facial expression recognition," in *2019 IEEE Visual Communications and Image Processing (VCIP)* (Sydney, NSW: IEEE), 1–4. doi: 10.1109/VCIP47243.2019.8965826

Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S., and FaceSoft, I. (2020). "Retinaface: Single-shot multi-level face localisation in the wild," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA: IEEE). doi: 10.1109/CVPR42600.2020.00525

Ding, H., Zhou, P., and Chellappa, R. (2020). "Occlusion-adaptive deep network for robust facial expression recognition," in *2020 IEEE International Joint Conference on Biometrics (IJCB)* 1–9 (Houston, TX: IEEE). doi: 10.1109/IJCB48548.2020.9304923

Farzaneh, A. H., and Qi, X. (2021). Facial expression recognition in the wild via deep attentive center loss," in 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 2401–2410 (Waikoloa, HI: IEEE). doi: 10.1109/WACV48630.2021.00245

Gao, S., Cheng, M.-M., Zhao, K., Zhang, X., Yang, M.-H., and Torr, P. H. S. (2019). Res2net: A new multi-scale backbone architecture. *IEEE Trans. Patt. Analy. Mach. Intell.* 43, 652–662. doi: 10.1109/TPAMI.2019.2938758

He, K., Zhang, X., Ren, S., and Sun, J. (2015). "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV: IEEE), 770–778. doi: 10.1109/CVPR.2016.90

Kim, J. H., Poulose, A., and Han, D. S. (2021). The extensive usage of the facial image threshing machine for facial emotion recognition performance. *Sensors.* 21, 2026. doi: 10.3390/s21062026

Li, H., Wang, N., Ding, X., Yang, X., and Gao, X. (2021). Adaptively learning facial expression representation via c-f labels and distillation. *IEEE Trans. Image Proc.* 30, 2016–2028. doi: 10.1109/TIP.2021.3049955

Li, S., and Deng, W. (2019). Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Trans. Image Proc.* 28, 356–370. doi: 10.1109/TIP.2018.2868382

Li, Y., Zeng, J., Shan, S., and Chen, X. (2018). "Patch-gated cnn for occlusion-aware facial expression recognition," in *2018 24th International Conference on Pattern Recognition (ICPR)* (Beijing: IEEE), 2209–2214. doi: 10.1109/ICPR.2018.8545853

Li, Y., Zeng, J., Shan, S., and Chen, X. (2019). Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Trans. Image Proc.* 28, 2439–2450. doi: 10.1109/TIP.2018.2886767

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J. M., Ambadar, Z., and Matthews, I. (2010). "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops* (San Francisco, CA: IEEE), 94–101. doi: 10.1109/CVPRW.2010.5543262

Ma, R., and Zhang, R. (2023). Facial expression recognition method based on PSA-YOLO network. *Front. Neurorob.* 16, 1057983. doi: 10.3389/fnbot.2022.1057983

Marrero-Fernández, P. D., Guerrero-Pe na, F. A., Tsang, I. R., and Cunha, A. (2019). "Feratt: Facial expression recognition with attention net," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (Long Beach, CA: IEEE), 837–846. doi: 10.1109/CVPRW.2019.00112

Pan, B., Wang, S., and Xia, B. (2019). "Occluded facial expression recognition enhanced through privileged information," in *Proceedings of the 27th ACM International Conference on Multimedia* (New York, NY: ACM), 566–573. doi: 10.1145/3343031.3351049

Poulose, A., Kim, J. H., and Han, D. S. (2021a). "Feature vector extraction technique for facial emotion recognition using facial landmarks," in *2021 International Conference on Information and Communication Technology Convergence (ICTC)* (Jeju Island: IEEE), 1072–1076. doi: 10.1109/ICTC52510.2021.9620798

Poulose, A., Reddy, C. S., Kim, J. H., and Han, D. S. (2021b). "Foreground extraction based facial emotion recognition using deep learning xception model," in *2021 Twelfth International Conference on Ubiquitous and Future Networks (ICUFN)* (Jeju Island: IEEE), 356–360. doi: 10.1109/ICUFN49451.2021.9528706

Proverbio, A. M., and Cerri, A. (2022). The recognition of facial expressions under surgical masks: The primacy of anger. *Front. Neurorob.* 16, 864490. doi: 10.3389/fnins.2022.864490

Pu, T., Chen, T., Xie, Y., Wu, H., and Lin, L. (2020). "Au-expression knowledge constrained representation learning for facial expression recognition," in *2021 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE), 11154–11161. doi: 10.1109/ICRA48506.2021.9561252

Sandler, M., Howard, A. G., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). "Mobilenetv2: Inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 4510–4520. doi: 10.1109/CVPR.2018.00474

She, J., Hu, Y., Shi, H., Wang, J., Shen, Q., and Mei, T. (2021). "Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Nashville, TN: IEEE), 6244–6253. doi: 10.1109/CVPR46437.2021.00618

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv*:1409.1556. doi: 10.48550/arXiv.1409.1556

Valstar, M. F., and Pantic, M. (2010). Induced disgust, happiness and surprise : an addition to the mmi facial expression database," in *Proceedings of the 3rd International Workshop on EMOTION (satellite of LREC)*: *Corpora for Research on Emotion and Affect*.

Wang, C., Wang, S., and Liang, G. (2019). "Identity- and pose-robust facial expression recognition through adversarial feature learning," in *Proceedings of the 27th ACM International Conference on Multimedia* (New York, NY: ACM), 238–246. doi: 10.1145/3343031.3350872

Wang, K., Peng, X., Yang, J., Lu, S., and Qiao, Y. (2020). "Suppressing uncertainties for large-scale facial expression recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA: IEEE), 6896–6905. doi: 10.1109/CVPR42600.2020.00693

Wang, K., Peng, X., Yang, J., Meng, D., and Qiao, Y. (2019). Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Trans. Image Proc.* 29, 4057–4069. doi: 10.1109/TIP.2019.2956143

Wang, X., Huang, J., Zhu, J., Yang, M., and Yang, F. (2018). "Facial expression recognition with deep learning," in *International Conference on Internet Multimedia Computing and Service* (ACM), 1–4. doi: 10.1145/3240876.3240908

Wen, Z., Lin, W.-L., Wang, T., and Xu, G. (2021). Distract your attention: Multi-head cross attention network for facial expression recognition. *Biomimetics* 8, 199. doi: 10.3390/biomimetics8020199

Xia, B., and Wang, S. (2020). "Occluded facial expression recognition with step-wise assistance from unpaired non-occluded images," in *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA: ACM), 2927–2935. doi: 10.1145/3394171.341 3773

Xue, F., Wang, Q., and Guo, G. (2021). "Transfer: Learning relation-aware facial expression representations with transformers," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (Montreal, QC: IEEE), 3581–3590. doi: 10.1109/ICCV48922.2021.0 0358

Xue, F., Wang, Q., Tan, Z., Ma, Z., and Guo, G. (2022). Vision transformer with attentive pooling for robust facial expression recognition. *IEEE Trans. Affec. Comput.* doi: 10.1109/TAFFC.2022.3226473

Zeng, J., Shan, S., and Chen, X. (2018). "Facial expression recognition with inconsistently annotated datasets," in *European Conference on Computer Vision* (Springer Nature Switzerland), 227–243. doi: 10.1007/978-3-030-01261-8_14

Zhang, F., Zhang, T., rong Mao, Q., and Xu, C. (2018). "Joint pose and expression modeling for facial expression recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 3359–3368. doi: 10.1109/CVPR.2018.0 0354

Zhang, Q.-L., and Yang, Y. (2021). "Sa-net: Shuffle attention for deep convolutional neural networks," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Toronto, ON: IEEE), 2235–2239. doi: 10.1109/ICASSP39728.2021.941 4568

Zhang, Y., Wang, C., and Deng, W. (2021). "Relative uncertainty learning for facial expression recognition," in *Neural Information Processing Systems*, eds. M., Ranzato, A., Beygelzimer, Y., Dauphin, P., Liang, J. W., Vaughan (Red Hook, NY: Curran Associates, Inc.), 17616–17627.

Zhao, G., Huang, X., Taini, M., Li, S., and Pietikäinen, M. (2011). Facial expression recognition from near-infrared videos. *Image Vision Comput.* 29, 607–619. doi: 10.1016/j.imavis.2011.07.002

Zhao, Z., Liu, Q., and Wang, S. (2021). Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Trans. Image Proc.* 30, 6544–6556. doi: 10.1109/TIP.2021.3093397

Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., and Torralba, A. (2015). "Learning deep features for discriminative localization," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV: IEEE), 2921–2929. doi: 10.1109/CVPR.2016.319

TYPE Original Research
PUBLISHED 24 August 2023
DOI 10.3389/fnbot.2023.1214203

OPEN ACCESS

EDITED BY
Long Jin,
Lanzhou University, China

REVIEWED BY
Alwin Poulose,
Indian Institute of Science Education and Research, Thiruvananthapuram, India
Guanqiu Qi,
Buffalo State College, United States

*CORRESPONDENCE
Yanyun Fu
✉ fyyun163@163.com

RECEIVED 29 April 2023
ACCEPTED 01 August 2023
PUBLISHED 24 August 2023

CITATION
Gan Y, Fu Y, Wang D and Li Y (2023) A novel approach to attention mechanism using kernel functions: Kerformer.
Front. Neurorobot. 17:1214203.
doi: 10.3389/fnbot.2023.1214203

A novel approach to attention mechanism using kernel functions: Kerformer

Yao Gan[1], Yanyun Fu[2]*, Deyong Wang[3] and Yongming Li[1]

[1]Information Science and Engineering Department, Xinjiang University, Ürümqi, China, [2]Beijing Academy of Science and Technology, Beijing, China, [3]Key Laboratory of Big Data of Xinjiang Social Security Risk Prevention and Control, Xinjiang Lianhai INA-INT Information Technology Ltd., Ürümqi, Xinjiang, China

Artificial Intelligence (AI) is driving advancements across various fields by simulating and enhancing human intelligence. In Natural Language Processing (NLP), transformer models like the Kerformer, a linear transformer based on a kernel approach, have garnered success. However, traditional attention mechanisms in these models have quadratic calculation costs linked to input sequence lengths, hampering efficiency in tasks with extended orders. To tackle this, Kerformer introduces a nonlinear reweighting mechanism, transforming maximum attention into feature-based dot product attention. By exploiting the non-negativity and non-linear weighting traits of softmax computation, separate non-negativity operations for *Query*(Q) and *Key*(K) computations are performed. The inclusion of the SE Block further enhances model performance. Kerformer significantly reduces attention matrix time complexity from $O(N^2)$ to $O(N)$, with N representing sequence length. This transformation results in remarkable efficiency and scalability gains, especially for prolonged tasks. Experimental results demonstrate Kerformer's superiority in terms of time and memory consumption, yielding higher average accuracy (83.39%) in NLP and vision tasks. In tasks with long sequences, Kerformer achieves an average accuracy of 58.94% and exhibits superior efficiency and convergence speed in visual tasks. This model thus offers a promising solution to the limitations posed by conventional attention mechanisms in handling lengthy tasks.

KEYWORDS

linear attention, kernel method, transformer, SE Block, self-attention

1. Introduction

The Transformer model and its variants have emerged as state-of-the-art approaches in various Artificial Intelligence (AI) tasks, including natural language processing (Devlin et al., 2018), computer vision (Carion et al., 2020; Dosovitskiy et al., 2020), and audio processing (Baevski et al., 2020), demonstrating impressive performance across a wide range of benchmarks. As evident from the Transformer model and its variants, researchers are continually exploring new methods and extensions to tackle challenges in different AI tasks, leading to remarkable achievements. For instance, in the field of speech emotion recognition, some works (Kakuba et al., 2022a,b) have made improvements to attention mechanisms, highlighting the widespread application and significance of Transformers and their extensions in diverse domains.

The core component of the Transformer is its attention mechanism, which efficiently encodes contextual information by modeling correlations between different positions in the input sequence. However, the original self-attention mechanism in the Transformer model, relying on dot product similarity, has limitations in modeling complex and non-linear relationships among tokens, and exhibits quadratic computational complexity concerning sequence length. Consequently, traditional Transformer models encounter challenges in handling long sequence data, particularly in terms of computational complexity and position information processing. Our approach aims to address this by reducing the time complexity of the attention matrix while maintaining accuracy in processing NLP tasks.

To overcome these challenges, researchers have proposed various extensions, including low-rank approximations, sparse patterns, and locality-sensitive hashing. Nevertheless, these methods still rely on dot product similarity and may not adequately capture diverse relationships among tokens. Recently, kernel methods have been introduced to enhance Transformer efficiency, allowing clever mathematical re-writing of the self-attention mechanism to avoid explicit computation of the N × N matrix.

In this paper, we propose a novel self-attention mechanism called Kerformer, which utilizes kernel functions to redefine the attention mechanism and extract richer positional information through reweighting. We conducted experiments on NLP and CV tasks, showing that Kerformer outperforms the original self-attention mechanism and other extensions in terms of accuracy and computational efficiency. Additionally, we performed an ablation study to analyze the impact of different kernel functions and reweighting positions on Kerformer's performance.

In comparison to state-of-the-art methods in self-attention and transformer architectures, our proposed Kerformer introduces a novel and efficient approach to self-attention computation. While previous works, such as Linformer (Wang et al., 2020), Reformer (Kitaev et al., 2020), DCT-Former (Scribano et al., 2023), LISA (Wu et al., 2021), and Bernoulli sampling attention mechanism (Zeng et al., 2021), have made significant strides in reducing computational costs and improving efficiency, they still rely on dot product similarity and may have limitations on sequence length and global dependencies.

In contrast, Kerformer leverages kernel methods to redefine the attention mechanism, enabling the capture of more complex and non-linear relationships among input tokens. By applying a kernel function and SE Block module to the concatenation of query and key vectors, Kerformer computes attention weights using the resulting kernel matrix, thereby modeling various types of relationships with enhanced expressiveness.

Moreover, our Kerformer introduces reweighting mechanisms that extract richer positional information, addressing challenges in long sequence processing and enhancing computational efficiency. This combination of kernel-based self-attention and reweighting sets Kerformer apart from existing approaches, making it a promising extension to the transformer architecture.

In the upcoming sections, we analyze existing self-attention methods and their limitations. We introduce the Kerformer model, discussing its novel kernel-based self-attention and reweighting mechanisms. We present experimental results and compare Kerformer with state-of-the-art methods on NLP and CV tasks. Finally, we discuss implications and conclusions in self-attention modeling.

In summary, our study introduces a novel self-attention mechanism, Kerformer, which utilizes compute kernels and reweighting techniques to capture intricate and diverse token interactions, while effectively addressing the computational complexity associated with long sequence tasks. By reducing the attention matrix complexity without compromising accuracy, Kerformer demonstrates its efficacy in various NLP and CV applications. Our research findings contribute to the advancement of more expressive and efficient self-attention mechanisms.

2. Related work

Self-attention has become a fundamental building block of modern neural architectures in natural language processing and computer vision. The original transformer architecture introduced by Vaswani et al. (2017) utilized self-attention as a key component to compute the representation of each input token. Since then, numerous variants of the transformer architecture have been proposed to overcome various limitations, such as the lack of position information and the quadratic complexity with respect to the sequence length.

Efforts have been made to improve the efficiency of self-attention, with several methods proposed to reduce computation costs. These include the Linformer (Wang et al., 2020), which approximates the self-attention matrix with a low-rank matrix, and the Reformer (Kitaev et al., 2020), which introduces locality-sensitive hashing to accelerate self-attention computation. DCT-Former (Scribano et al., 2023) achieves efficient self attention computation by introducing discrete cosine transform as a frequency domain based conversion method. By calculating attention weights in the frequency domain, DCT-Former can significantly reduce computational complexity while maintaining high performance, improving the efficiency and scalability of the model. LISA (Wu et al., 2021) utilizes a codeword histogram technique to achieve linear-time complexity for self-attention computation. By representing tokens as codewords and constructing histograms based on their frequencies, the model efficiently captures token interactions and calculates attention weights. This approach reduces the computational overhead associated with traditional self-attention mechanisms, making it suitable for large-scale recommendation tasks. A Bernoulli sampling attention mechanism (Zeng et al., 2021) based on locally sensitive hashing (LSH) approximates the calculation of self attention weights through random sampling, thereby reducing computational complexity to a linear level. The Bernoulli sampling method can significantly reduce the time and space overhead of self attention computation while maintaining good performance. However, the above methods often have limitations on the length of the sequence and limit the global dependencies of the sequence.

In addition, there are attempts to extend self-attention beyond its original formulation. For example, the Sparse Transformer (Child et al., 2019; Beltagy et al., 2020; Zaheer et al., 2020) introduces sparsity patterns to reduce computational costs. The Performer (Choromanski et al., 2020) uses an approximation of the softmax function to compute self-attention more efficiently.

Moreover, Katharopoulos et al. (2020) reformulated the attention mechanism in the autoregressive Transformer model to use sequential computation, thereby reducing computation time and storage requirements. Nyströmformer (Xiong et al., 2021) proposed a method based on Nyström approximation, which approximates the calculation of self attention weight by decomposing the self attention Matrix decomposition into the product of low rank matrix. Nevertheless, these approaches may also exhibit certain limitations, including elevated memory usage, potential degradation of model accuracy, or approximation errors.

Recently, kernel-based methods have emerged as a promising extension of self-attention. Kernel methods replaces the dot-product similarity used in self-attention with a kernel function, allowing it to capture more complex interactions between input tokens and enabling the use of more powerful kernel functions to model various types of relationships. This method allows iterative implementation, which significantly accelerates Transformer and reveals their relationship with recurrent neural networks. The Kernel methods mechanism has been successfully applied to various tasks, such as text classification and image classification. Skyformer (Chen et al., 2021) proposes a novel approach that employs a Gaussian kernel and the Nyström method to approximate self-attention, thereby reducing computational complexity while maintaining accuracy. This work shows promising results on several natural language processing tasks, including text classification and machine translation. Kernel self-attention (Rymarczyk et al., 2021) proposes a novel approach for weakly-supervised image classification by combining kernel self-attention with deep multiple instance learning. The method uses a kernel function to capture complex interactions between image regions and enable more powerful modeling of relationships.

Several modifications to attention have been proposed by researchers, including the use of softmax to operate Q and K matrices separately (Bhandare et al., 2019), and the decomposition of attention into kernel functions, with Q and K matrices operated on using the *elu* and *relu* functions, respectively (Katharopoulos et al., 2020; Qin et al., 2022). These modifications reduce the complexity of attention from $O(N^2)$ to $O(N)$, which is beneficial for large-scale models.

In comparison to the state-of-the-art methods in self-attention and transformer architectures, our proposed Kerformer introduces a novel and efficient approach to self-attention computation. While previous works, such as Linformer, Reformer, DCT-Former, LISA, and Bernoulli sampling attention mechanism, have made significant strides in reducing computational costs and improving efficiency, they still rely on dot product similarity and may have limitations on sequence length and global dependencies. In contrast, Kerformer leverages kernel methods to redefine the attention mechanism, enabling the capture of more complex and non-linear relationships among input tokens. By applying a kernel function and SE Block module to the concatenation of query and key vectors, Kerformer computes attention weights using the resulting kernel matrix, thereby modeling various types of relationships with enhanced expressiveness.

Moreover, our Kerformer introduces reweighting mechanisms that extract richer positional information, addressing challenges in long sequence processing and enhancing computational efficiency. This combination of kernel-based self-attention and reweighting sets Kerformer apart from existing approaches, making it a promising extension to the transformer architecture.

In conclusion, self-attention has undergone significant developments since its introduction in the original transformer architecture, with research focusing on improving its efficiency, scalability, and expressiveness. Kernel methods is a recent extension that shows promise in modeling complex relationships between input tokens, and several modifications have been proposed to enhance its performance. The Kerformer proposed in this study addresses the existing research gap by introducing kernel functions and reweighting mechanisms, effectively tackling challenges in long sequence processing and enhancing computational efficiency. The main idea of Kerformer is to change the order of operations of matrices according to the union law of matrices, so as to linearize the attention. When linearizing the attention, we first activate the Q and K matrices through the activation function to ensure the non-negativity of the attention matrix, and then reweight the K matrix through the SE-K module to achieve the redistribution of attention, so as to improve the performance of the model.

3. Methodology

In this section, we propose a novel linear Transformer model called **Kerformer**. We introduce a decomposable linear attention mechanism that replaces traditional softmax attention, resulting in improved time and memory complexity. Our method is also applicable to casual attention. The **Kerformer** model also employs different activation functions for Q and K, and combined with SE Block to reweight the activated K, which contributes to its faster computing speed and better performance.

3.1. Transformer

Given an input sequence x of length N and feature dimension d, we represent it as $x \in \mathbb{R}^{N\times d}$. The Transformer model can be formulated as Eq. 1.

$$T(x) = F((A(x) + x) \tag{1}$$

In the Transformer model, the F implementation typically corresponds to a feedforward neural network that transforms the characteristics of each input. The attention function is denoted by A, and its time and memory complexity scales quadratically with respect to the input sequence length N.

The core idea of the attention mechanism is that the network should give different importance to different parts of the input data. When processing the input data, the network needs to assign different weights to different parts of the input in order to better capture the important information in the input data. This process of weight assignment is the attention mechanism.

In implementing the attention mechanism, two key components are usually used: *query*(Q), *key*(K), and *value*(V). A query is a vector in the network that represents the network's attention to the input data. Keys and values are vectors in the input data used to represent different parts of the input data. The attention mechanism achieves attention to the input data by computing the similarity between the query and the key and assigning weights to the values based on the similarity.

Regarding the attention function A, it consists of three essential components, including *query*(Q), *key*(K), and *value*(V). These components are computed from the input sequence x and three learnable matrices W_Q, W_K, and W_V, respectively, as follows: $Q = xW_Q, K = xW_K, V = xW_V$.

The final output $A = V'$ is obtained through a softmax function applied to QK^T line by line, which can be expressed as follows in Eq. 2.

$$A(x) = V' = softmax(\frac{QK^T}{\sqrt{D}})V \tag{2}$$

We can interpret Eq. 2 as a specific instance of the attention mechanism, where the softmax function is applied to calculate QK^T. In order to introduce a more generalized expression of attention, we can use V_i to represent the i-th row of a matrix $V(V \in \mathbb{R}^{N\times d})$. The equation of the generalized attention mechanism is shown below as Eq. 3. Similar derivations have been done in these works (Qin et al., 2022).

$$V_i' = \sum_{j=1}^{N} \frac{sim(Q_i, K_j)}{\sum_{j=1}^{N} sim(Q_i, K_j)} V_j \tag{3}$$

It should be noted that the function sim in Eq. 3 can be any correlation function that satisfies certain requirements, which will be explained later. If we choose $sim(Q, K) = e^{\frac{QK^T}{\sqrt{d}}}$, then Eq. 3 is equivalent to Eq. 2.

3.2. Linear attention

To maintain the linear computation budget, one feasible solution is to expand the *sim* function in the form of a kernel function, as shown in Eq. 4.

$$sim(q_i, k_j) = \phi(q_i)^T \varphi(k_j) \tag{4}$$

In Eq. 3, ϕ and φ are kernel functions used for the nonlinear mapping of queries and keys. We can rewrite Eq. 3 as a kernel function, as shown in Eq. 5.

$$V_i = \frac{\sum_{j=1}^{N}(\phi(Q_i)\varphi(K_j)^T)V_j}{\sum_{j=1}^{N}(\phi(Q_i)\varphi(K_j)^T)} \tag{5}$$

Then, the attention operation under linear complexity can be realized through the multiplication combination law of matrix, as shown in Eq. 6.

$$V_i = \frac{\phi(Q_i)\sum_{j=1}^{N}\varphi(K_j)^T V_j}{\phi(Q_i)\sum_{j=1}^{N}\varphi(K_j)^T} \tag{6}$$

Note that in Eq. 4, the functions ϕ and φ are applied row by row to the matrices Q and K. By using the associative law of multiplication, $QK^T \in \mathbb{R}^{N\times N}$ is calculated as $\varphi(K)^T V \in \mathbb{R}^{d\times d}$. The result is then left multiplied by $\phi(Q) \in \mathbb{R}^{N\times d}$, which represents the attention weights. This computation mode achieves a complexity of $O(Nd^2)$ for the attention mechanism. However, for long sequences where $d \ll N$, the complexity can be considered as $O(N)$, greatly reducing the overhead. This is illustrated in Figure 1.

3.3. Kerformer

The softmax operation applied in the attention mechanism is used to normalize the query and key matrices. However, there is no clear explanation for why the softmax operation is effective, and it is more of an empirical observation that leads to good model performance. Our aim is to enhance the attention mechanism by using the kernel form. Specifically, we want to generalize the attention mechanism using the kernel function and provide a theoretical foundation for the application of different operations in the attention mechanism. This will help us better understand the working principles of the attention mechanism and improve its performance.

Cosformer (Qin et al., 2022) discussed that the choice of ϕ and φ functions is crucial for the performance of attention mechanisms in kernel form. They proposed two empirical constraints that may play a significant role in achieving better performance:

(i) Non-negative constraint on the attention matrix to ensure that the attention weights are always positive and the attention is focused only on relevant features.

(ii) A nonlinear weighted scheme to focus attention on specific regions of the matrix distribution, which can capture more complex and subtle patterns.

It is worth noting that similar kernel function methods have been used to modify the attention mechanism in the works of Angelos and Qin et al. These works always choose the same activation function for both the ϕ and φ functions. We decided to choose different ϕ and φ functions to enhance the model's global learning ability and generalization ability.

To ensure the two constraints mentioned above, we use sigmoid activation function for $\phi(Q)$ and softmax activation function for $\varphi(K)$ instead of the original $softmax(QK^T)$ in our work. Thus, we define our functions as shown in Eq. 7 and Eq. 8.

$$\phi(x) = sigmoid(x) \tag{7}$$

$$\varphi(x) = softmax(x) \tag{8}$$

We substitute Eqs 7 and 8 into Eq. 6 to obtain Eq. 9, as follows:

$$V_i = \frac{sigmoid(Q_i)\sum_{j=1}^{N} softmax(K_j)^T V_j}{sigmoid(Q_i)\sum_{j=1}^{N} softmax(K_j)^T} \tag{9}$$

The system block diagram of Kerformer is shown in Figure 2.

3.4. Interpretation of Kerformer

Previous works, such as Katharopoulos et al. (2020) and Qin et al. (2022), have also rewritten self-attention in kernel form, but they have used the same function to transform both the Q

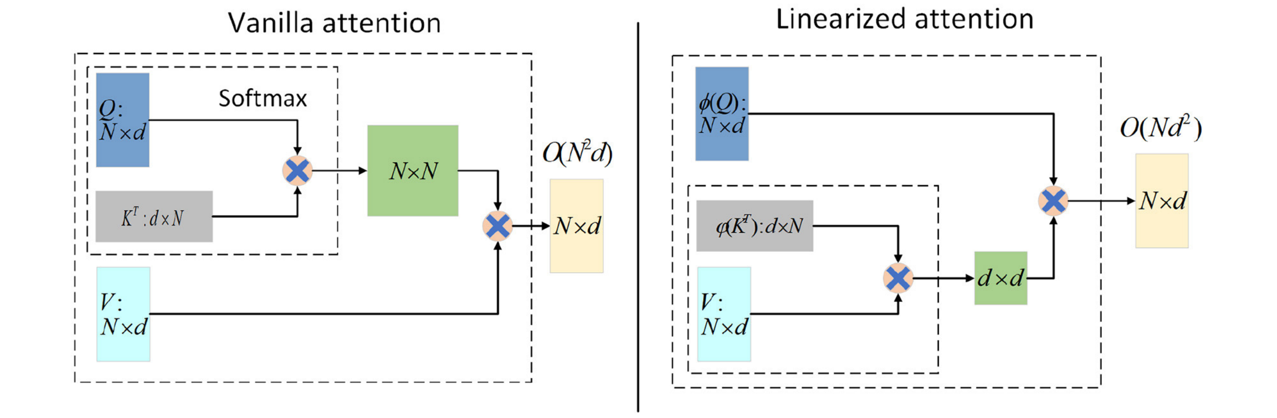


FIGURE 1
Illustration of the computations for Vanilla attention **(left)** and Linearized attention **(right)**. For input, the input length is N and the feature dimension is d. ϕ and φ represent the kernel function form for processing Q and K. Generally speaking, $d \ll N$, Linearized attention can be approximately regarded as the time and memory complexity of $O(N)$.

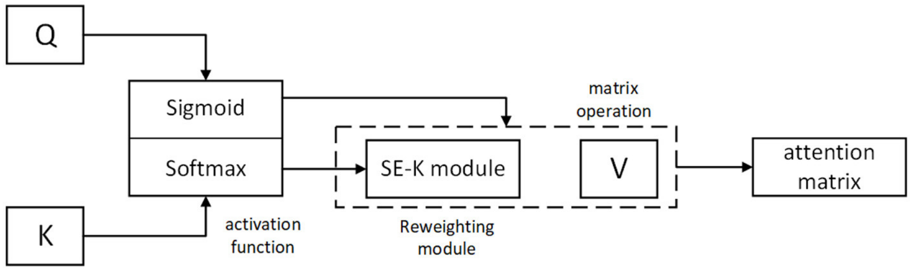


FIGURE 2
System block diagram of our approach Kerformer and workflow representation.

and K matrices. The possible reason for this is that if different transformations are applied to the Q and K matrices, the relative positional relationship between them may be disrupted. This could lead to inaccurate score calculations and negatively affect the performance of the model.

However, Efficient attention (Shen et al., 2021) provided a new explanation for their proposed linear attention, which is different from self attention. They explained that linear attention does not generate attention maps for each position, and each $(K_j)^T$ is a global attention map that does not correspond to any position. Based on this explanation, we aim to introduce different functions for Q and K without disturbing the attention mechanism as much as possible, which may bring improvements to the model.

The explanation provided by Efficient attention (Shen et al., 2021) regarding linear attention inspired our work to introduce different functions for Q and K matrices. This would allow us to explore new explanations and extensions to the attention mechanism.

Our approach includes introducing different nonlinear mappings for Q and K matrices. We use the sigmoid operation on Q to limit its range between 0 and 1, mapping each element to a probability distribution. Similarly, we apply the softmax operation on K to also map each element to a probability distribution. This introduces more nonlinearity to the model, making it better suited to fit the data.

Furthermore, the model is forced to learn different information due to the effects of these operations. The sigmoid operation allows the model to focus more on keys that are similar to the query, while the softmax operation enables the model to focus more on elements with higher probabilities in the values. This combination allows the model to learn better in different directions.

Lastly, the use of the smooth sigmoid and softmax operations makes the model more robust to data disturbance or noise, reducing the risk of overfitting. Overall, our approach introduces new insights into the attention mechanism and improves the model's performance.

3.5. Reweighting of attention

The above explanation highlights the difference between linear attention and self-attention, with linear attention not generating attention maps for each position. Given this difference, we aim to introduce the SE module to perform re-weighting of the K matrix along the N dimension. The goal is to extract different features by using different functions for Q and K without disturbing the attention mechanism as much as possible, which could lead to improvements in the performance of the model. By using the SE module, we can dynamically recalibrate the feature maps of K based

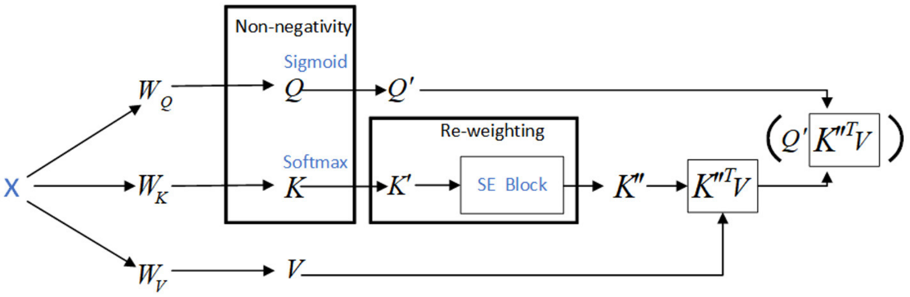


FIGURE 3
Use the activation functions *Sigmoid* and *Softmax* to activate the *Q* and *K* matrices respectively.

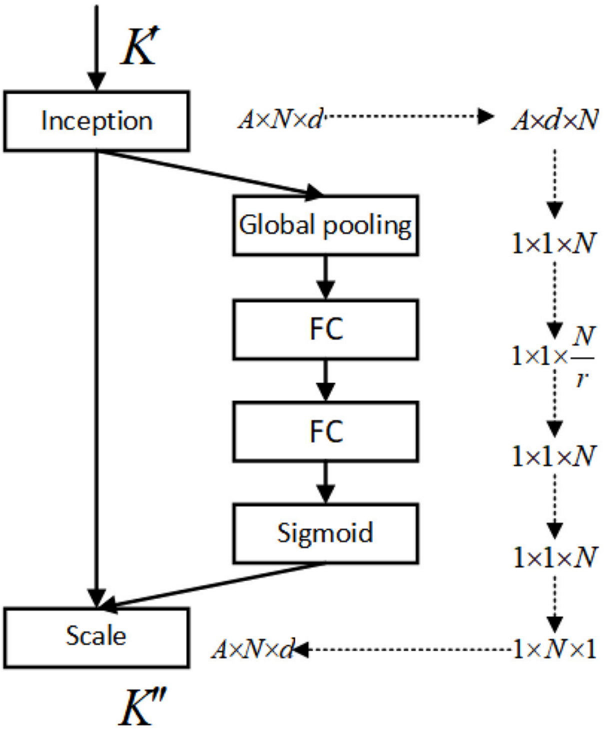


FIGURE 4
The structure of the SE-K module is shown, and the K' matrix after activation by *Softmax* is reweighted.

on their importance, thus improving the model's ability to extract meaningful information from the input data.

In order to adapt to the reweighting of the *K* matrix, we slightly modified the SE module and referred to it as the SE-K module. As mentioned earlier, the *K* matrix itself already possesses non-negative values, we remove the ReLU activation function from the SE module. The SE-K module is a modified version of the SE module that takes into account the non-negativity of the *K* matrix.

In this section, we will describe how we incorporate the SE-K module into the *K* matrix of the attention mechanism. Specifically, we apply the SE-K module to the N dimension of the *K* matrix, where K has a dimension of N x d.

The SE module is a simple yet effective mechanism that is widely used to enhance the representational power of neural networks. It selectively recalibrates the feature map by using the global information of the feature map. In our method, we use the SE-K module to recalibrate the K matrix, thereby improving its feature extraction ability.

To apply the SE-K module to the *K* matrix, we first perform a global pooling operation on the *K* matrix along the N dimension, resulting in a feature vector. This feature vector is then passed through two fully connected layers, which are followed by a sigmoid activation function. The output of the sigmoid function is a set of N-dimensional attention weights, which are used to weight the *K* matrix along the N dimension. Finally, the weighted *K* matrix is fed into the attention mechanism. The operation to activate the *Q* and *K* matrices is shown in Figure 3, and the network structure of the SE-K module involved is shown in Figure 4.

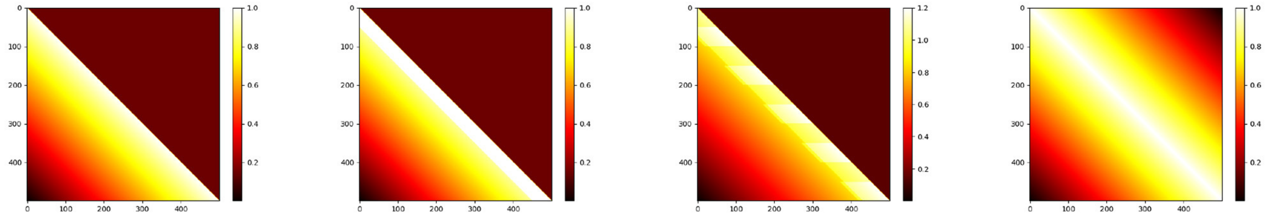
For NLP tasks, Kerformer places more weight on neighboring tokens, thus enhancing locality. The weight distribution is shown in the Figure 5. By using the SE-K module, we can effectively learn the importance of different features in the *K* matrix, which can significantly improve the performance of the attention mechanism. Additionally, the SE-K module has a relatively small computational cost, which makes it easy to incorporate into existing neural network architectures.

Overall, our method of applying the SE-K module to the *K* matrix has shown promising results in various tasks, demonstrating its effectiveness in improving the feature extraction ability of the attention mechanism.

Our research method is based on the activation function and the reweighting mechanism. The activation function is to perform a non-negativity operation on the matrix to satisfy the requirement of non-negativity of the attention matrix, while the reweighting operation is to redistribute the attention weights to achieve the effect that the local influence on the nearby attention is greater. These two operations can better satisfy the attention relationship between different parts to obtain the final attention matrix. For data collection we use all the data sets that are now publicly available and conduct our experiments on these publicly available datasets.

4. Simulation experiments

In this section, we present an evaluation of our proposed method, Kerformer, through simulation experiments. The simulation experiment focuses on a mathematical evaluation of



(1) Vanilla Transformer (2) Kerformer (3) Kerformer w/o re-weighting matrix (4) Kerformer re-weighting matrix

FIGURE 5
(1): Attention matrix of vanilla transformer. (2): Attention matrix of Kerformer. (3): Attention matrix of Kerformer without re-weighting. (4): Visualization of the re-weighting matrix.

Kerformer. We compare our model with four baselines, Vanilla attention (Vaswani et al., 2017), Efficient attention (Shen et al., 2021), Linear-Elu (Katharopoulos et al., 2020), and Performer (Choromanski et al., 2020), to demonstrate the superiority of our approach in terms of model running memory, running time. All experiments were conducted using Matlab R2020a.

4.1. Comparison of time costs in simulation experiments

This experiment fixes the number of input matrices as 1 and the attention head dimension as 64, and compares the running time of each method by changing the sequence length size N of input x. The specific results can be seen in Table 1, with time units in seconds.

From the experimental results in Table 1, we can see that four other methods have a greater advantage over the Vanilla attention method in terms of the time cost of attention matrix computation, especially Vanilla attention has experienced memory overflow when the input sequence length N is large. In addition, our proposed method usually outperforms other methods with shorter computation time when the length of the input sequence N is below the million level. In practice, the model input length N is always below the million level. That is, our proposed method outperforms other methods in use.

From the experimental results in Table 2, it can be seen that four other methods have time cost advantages over Vanilla attention to different ranges of Q, K, and V values. Cosformer has more time cost advantage in computing Attention when the value range is $[-10,10]$, while our method has a shorter running time compared to the other three methods for the range of values of Q, K, and V below $[-10,10]$, which fully illustrates the advantage of our method in terms of time cost.

4.2. Comparison of memory costs in simulation experiments

The experimental results in Table 3 show that the other four methods have a smaller memory consumption compared to the Vanilla attention method in the computation of the attention matrix. According to our empirical observation, the value range of Q, K, and V matrices input into the attention mechanism is mostly between $[-4,4]$. Our method has a memory cost advantage in the range of $[-2,2]$ and $[-4,4]$, which indicates that our method can achieve a low memory cost in the normal range of values, which can be attributed to the fact that our method uses different activation functions for Q and K, which can improve the computational speed and generalization ability of the model.

5. NLP task

We empirically validate the effectiveness of our proposed Kerformer method in multiple aspects. Firstly, we examine its generalization capability on downstream tasks by comparing it with other existing transformer variants. Then, we conduct a comparison with other Long-range arena benchmark transformer variants to assess its ability to model long-range dependencies and to perform a thorough analysis of model efficiency.

5.1. Downstream fine-tuning tasks

First, we performed the Kerformer model and the remaining five models [Performer (Choromanski et al., 2020), Reformer (Kitaev et al., 2020), and Liner Trans (Katharopoulos et al., 2020), Longformer (Beltagy et al., 2020), RFA (Peng et al., 2021), and Dct-former (Scribano et al., 2023)] were compared in terms of accuracy. This was achieved by conducting comparative fine-tuning experiments on five datasets, including GLUE (QQP, SST-2, MNLI) (Wang et al., 2018), IMDB (Maas et al., 2011), and Amazon (Ni et al., 2019). In the experiments, pre-trained models are used and fine-tuned in the downstream text classification task, and the results are shown in Table 4. From Table 4, we can see that Kerformer fetches the best accuracy in addition to the baseline (Liu et al., 2019) on the QQP, SST-2 and IMDB downstream text classification tasks. Although Dct-former and Longformer achieved better classification accuracy than Kerformer on MNLI and AMAZON tasks, respectively. It has higher computational complexity compared to our method. This is related to Kerformer's activation of Q and K matrices with activation functions and reweighting of K matrices respectively, where the activation functions can extract features in the matrices and reweighting can effectively reallocate attention to achieve the effect of expanding

TABLE 1 Comparison of the time required to run the five methods for different methods in different dimensions of the input *x*, *Q*, *K*, and *V* in the case of taking values in the range [−2,2].

Dimensional changes	Vanilla attention	Efficient attention	Linear-Elu	Performer	Kerformer (ours)
1*1,000*64	4.001 s	1.000 s	1.000 s	0.882 s	0.200 s
1*10,000*64	302.072 s	31.015 s	6.0121 s	6.112 s	5.852 s
1*100,000*64	OOM	87.024 s	51.014 s	55.514 s	44.011 s
1*1,000,000*64	OOM	967.22 s	506.134 s	505.514 s	521.144 s

TABLE 2 Comparison of the time required to run the five methods with different ranges of values for *Q*, *K*, and *V* for different methods with the dimension size of the input *x* of 1*10,000*64.

Range of values	Vanilla attention	Efficient attention	Linear-Elu	Performer	Kerformer (ours)
[−1,1]	335.075 s	34.007 s	7.001 s	6.854 s	6.001 s
[−2,2]	302.072 s	31.015 s	6.012 s	6.112 s	5.852 s
[−4,4]	1,003.233 s	35.008 s	5.025 s	6.012 s	5.006 s
[−6,6]	1,062.249 s	34.008 s	5.145 s	5.541 s	5.022 s
[−8,8]	1,032.248 s	35.993 s	6.004 s	6.125 s	5.952 s
[−10,10]	1,103.246 s	55.013 s	8.001 s	7.854 s	8.004 s

TABLE 3 Comparison of the memory requirements of the five methods running with different ranges of values for *Q*, *K*, and *V* for the input *x* with dimension size of 1*10,000*64.

Range of values	Vanilla attention	Efficient attention	Linear-Elu	Performer	Kerformer (ours)
[−1,1]	8,521 M	521 M	623 M	689 M	534 M
[−2,2]	11,001 M	585 M	678 M	702 M	578 M
[−4,4]	12,454 M	623 M	725 M	754 M	602 M
[−6,6]	14,845 M	685 M	775 M	801 M	692 M
[−8,8]	15,624 M	725 M	835 M	833 M	754 M
[−10,10]	16,104 M	785 M	877 M	892 M	802 M

local attention. The experimental result fully demonstrates the effectiveness of our proposed Kerformer model.

5.2. Long sequence experiment results

To assess the generalization performance of our proposed method Kerformer, we conducted training from scratch on the Long-range Arena benchmark 2020b. This benchmark is tailored for evaluating the performance of efficient transformers on long input sequences, making it an appropriate test platform for comparative analysis of different efficient transformer variants. We evaluated our approach on various tasks, including long sequence ListOps (Nangia and Bowman, 2018), byte-level text classification (Maas et al., 2011), document retrieval using ACL selection networks (Radev et al., 2013), and Pathfinder (Linsley et al., 2018). While comparing with our Kerformer model with Local Attention (Tay et al., 2020), Reformer (Kitaev et al., 2020), Performer (Choromanski et al., 2020), Longformer (Choromanski et al., 2020), Transformer (Vaswani et al., 2017), BigBird (Zaheer et al., 2020), and Dct-former (Scribano et al., 2023) models, the comparison results of the seven different models are shown in Table 5. As shown in Table 5, Kerformer obtained the best performance in ListOps, Document Retrieval, while Kerformer also achieved competitive results in the other two tasks, and finally Kerformer achieved the next best score in overall task average accuracy. This is a good indication of Kerformer's strength in the long-range arena.

5.3. Ablation experiments

To verify the effectiveness of our chosen activation function in combination with the SE-K module, we conducted ablation experiments on GLUE (QQP, SST-2) (Wang et al., 2018) and IMDB (Maas et al., 2011) in downstream fine-tuning tasks, ListOps (Nangia and Bowman, 2018) in Long sequence tasks, byte-level text classification (Maas et al., 2011) and document retrieval using ACL selection networks (Radev et al., 2013) were conducted for the ablation experiments, and the results of the experiments are shown in the following Table 6.

As shown in Table 6, Q + Softmax(K)+SE-K indicates that no activation operation is performed on the Q matrix, Sigmoid(Q) + K + SE-K indicates that no activation operation is performed

TABLE 4 Results of fine-tuning downstream tasks based on pretrained bidirectional models.

	QQP ↑	SST-2 ↑	MNLI ↑	IMDB ↑	AMAZON ↑	Avg ↑
Vanilla transformer	88.52	92.25	80.02	92.55	75.65	85.80
Performer	69.95	50.82	35.28	60.41	64.25	56.14
Reformer	63.12	50.66	35.35	49.88	64.32	52.67
Liner Trans	74.75	84.72	66.35	91.21	72.62	78.07
Longformer	85.55	88.56	77.27	91.07	**73.52**	83.13
RFA	75.32	76.44	57.71	78.86	68.08	71.28
Dct-former	85.56	86.89	**77.48**	89.68	72.12	80.19
Kerformer	**85.68**	**90.21**	76.32	**91.50**	73.24	**83.39**

Best results are shown in bold. Our proposed Kerformer shows superior performance compared to competing efficient transformers and is approaching vanilla transformers.

TABLE 5 Long-range arena benchmark test results.

Model	ListOps ↑	Text ↑	Retrieval ↑	Pathfinder ↑	Avg ↑
Local attention	15.67	52.87	53.40	66.59	47.13
Reformer	37.32	56.12	53.42	68.47	53.83
Performer	17.96	65.45	53.79	**77.08**	53.57
Longformer	35.65	62.79	56.83	69.69	56.24
Transformer	36.42	64.37	57.52	71.42	57.43
BigBird	36.11	64.08	59.31	74.79	58.57
Dct-former	36.55	**65.15**	59.55	75.56	**59.20**
Kerformer	**36.95**	64.32	**59.98**	74.52	58.94

The best results are shown in bold and the second best results are underlined. Kerformer obtained the best average score in four different tasks.

TABLE 6 Ablation experiments are performed for the SE Block in the downstream fine-tuning task and the long sequence task of the reweighting module.

Model structure	QQP	SST-2	IMDB	ListOps	Text	Retrieval
Q + Softmax(K) + SE-K	81.25	85.63	85.24	33.25	58.53	55.89
Sigmoid(Q) + K + SE-K	82.36	87.25	88.25	35.21	60.25	57.26
Sigmoid(Q) + Softmax(K)	81.26	85.09	85.18	32.23	57.87	56.31
Kerformer	85.68	90.21	91.50	36.95	63.32	59.98

on the K matrix, and Sigmoid(Q) + Softmax(K) indicates that no reweighting operation is performed. Based on the results of the ablation experiments, it can be seen that the activation of the Q and K matrices and the reweighting operation on the K matrix can effectively improve the performance of the model in the downstream fine-tuning task and the long-sequence task relative to other methods, and the effectiveness of our method is also demonstrated.

5.4. Efficiency comparison

In addition to comparing model performance, we also compared the computational speed of the different models. We compared the computational speed of Kerformer with other models [standard Transformer (Vaswani et al., 2017), Local Attention (Tay et al., 2020), Reformer (Kitaev et al., 2020), BigBird (Zaheer et al., 2020), Linear Trans (Katharopoulos et al., 2020), Performer (Choromanski et al., 2020), Longformer (Beltagy et al., 2020), and Dct-former (Scribano et al., 2023)], and the variable for comparison was the length of the input sequence, and the results of the experiments are shown in Table 7. We used byte-level text classification benchmarks to measure the computational speed of different models during training and inference for different sequence lengths (1k–4k).

Our method Kerformer achieves good training and inference speeds on sequence lengths 2K, 3K, and 4K, which illustrates the advantage of our method for speed computation on long sequence let tasks. This is because first the Q and K matrices are activated, then the K matrices are reweighted separately, and finally the order of computation of the self-attentive matrices can be exchanged using the union law of matrices so that the goal of linear complexity can be achieved. In conclusion, our model Kerformer achieves better overall efficiency compared to other linear variables, while maintaining excellent modeling and generalization capabilities.

TABLE 7 Speed comparison in training and inference for long-range arena benchmarks with different sequence lengths (1–4k).

	Inferrence speed (steps per second)↑				Train speed (steps per second)↑			
Model	1K	2K	3K	4K	1K	2K	3K	4K
Transformer	25.42	7.85	\	\	6.91	2.19	\	\
Local attention	57.69	33.21	23.32	17.80	13.42	6.61	4.35	3.10
Reformer	44.23	21.60	12.75	8.35	11.60	5.01	2.96	1.97
BigBird	20.92	11.53	8.14	6.12	6.50	3.21	2.09	1.55
Linear Trans	67.81	38.22	26.30	19.92	11.88	5.56	3.54	2.49
Performer	74.20	42.35	29.53	22.43	14.23	6.50	4.13	2.93
Longformer	23.02	6.33	\	\	4.42	1.31	\	\
Dct-former	56.21	34.21	22.85	20.51	11.58	5.95	3.92	2.32
Kerformer	57.42	33.15	21.45	17.13	11.34	5.58	3.57	2.55

If a method runs out of memory, we mark it with a backslash. The higher it is, the better it is.

6. Visual classification task

By incorporating distinct functions into the Q and K matrices, Kerformer is specifically designed to facilitate feature extraction at different levels, which is highly advantageous for visual classification tasks. The primary objective of our study is to showcase the superior performance of Kerformer in such tasks. To achieve this, we conducted comprehensive image classification experiments to rigorously evaluate the effectiveness and efficiency of Kerformer.

In order to assess the performance of Kerformer in image classification tasks, we applied it to the widely-used ViT-B/16 (Dosovitskiy et al., 2020) model and compared its accuracy with that of several baseline models, including Vanilla attention (Vaswani et al., 2017), Efficient attention (Shen et al., 2021), Linear-Elu (Katharopoulos et al., 2020), and Cosformer (Qin et al., 2022). To this end, we evaluated the models on four datasets: MNIST, CIFAR-10, CIFAR-100, and the flower dataset provided by TensorFlow.

The MNIST dataset consists of handwritten digital images, consisting of 60,000 training images and 10,000 test images, each representing a gray number from 0 to 9. Cifar-10 is a widely-used computer vision dataset for object recognition, comprising 60,000 RGB color images with dimensions of 32 × 32 pixels, distributed across 10 different classes. CIFAR-100 dataset contains 100 classes, grouped into 20 superclasses. Each image in CIFAR-100 is labeled with a "fine" class (specific class) and a "coarse" class (superclass). The flower dataset includes images of daisies and encompasses five flower types: "daisy," "dandelion," "rose," "sunflower," and "tulip."

Overall, our results suggest that Kerformer has strong feature extraction ability and outperforms the baseline models in terms of accuracy.

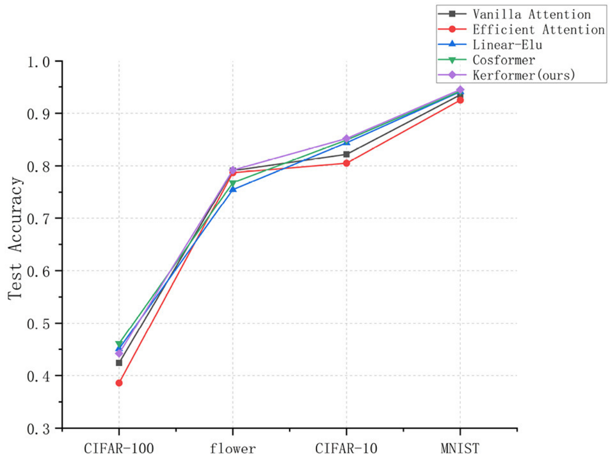


FIGURE 6
Experimental results of image classification accuracy measured by models using five methods (Vanilla attention, Efficient attention, Linear-Elu, Cosformer, and Kerformer) on different datasets.

6.1. Test accuracy

In this section, we performed accuracy tests on the image classification tasks using the aforementioned four datasets. For all datasets except the flower dataset, the experiments were conducted with the following settings: the images were resized to 224 × 224 pixels, Adam optimizer was employed, the learning rate was set to 0.0001, the loss function used was Cross Entropy, the batch size was set to 32, and the training was carried out for 180 epochs. The final test accuracy was computed by averaging the results of 10 test runs. Due to the limited size of the flower dataset, the experimental configuration differed in terms of a smaller batch size of 4, a reduced training epoch of 80, and the final test accuracy was determined by averaging the results of 10 test runs.

Based on the experimental results shown in Figure 6, it is evident that the Cosformer method can achieve the highest model accuracy for image classification on the CIFAR-100 dataset, whereas our proposed method can achieve the highest test accuracy for image classification on the MNIST, CIFAR-10, and flower datasets. In particular, our method can improve 3% points compared to Vanilla attention method on CIFAR-10 dataset, which is a better test for the model performance improvement of the original model. Our results suggest that our proposed improvement can

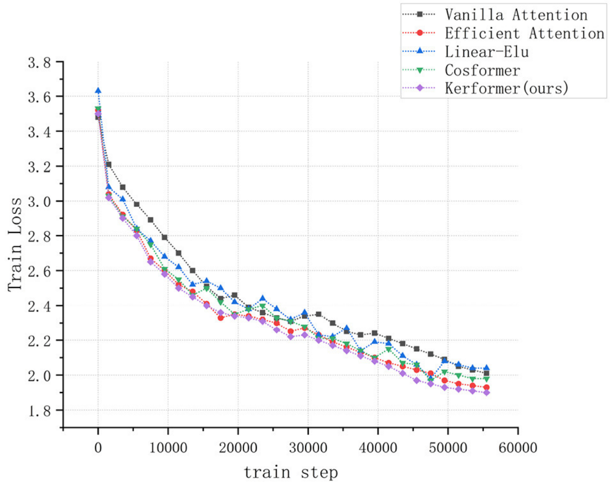


FIGURE 7
Comparison of convergence speed of ViT models using Vanilla attention, Efficient attention, Linear Elu, Cosformer and Kerformer when trained on the CIFAR-10 dataset.

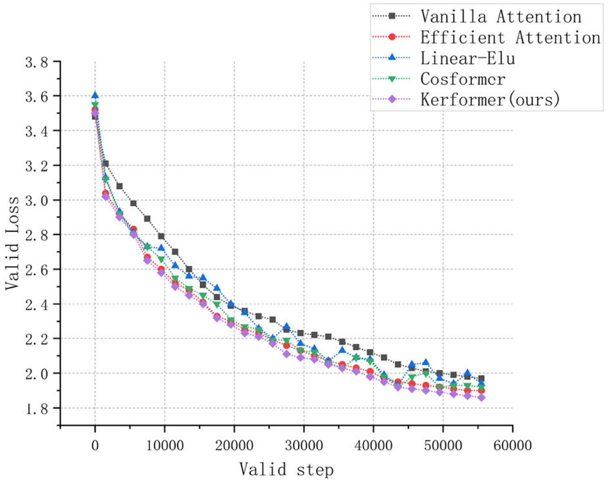


FIGURE 8
Comparison of convergence speed of ViT models with Vanilla attention, Efficient attention, Linear Elu, Cosformer and Kerformer when validated on CIFAR-10 dataset.

significantly enhance the performance of the model. In particular, this enhancement enables the model to more effectively utilize feature information from various locations, thereby improving its ability to extract essential features and ultimately increasing the classification accuracy of the model. This is due to the use of operations such as pooling in the SE-K module, which can perform better in image tasks because it is not limited by the global nature.

6.2. Convergence speed

In addition to evaluating the model performance and running cost, we also conducted experiments to measure the convergence speed of the ViT model during training and validation on the CIFAR-10 dataset using three methods: Vanilla attention (Vaswani et al., 2017), Efficient attention (Shen et al., 2021), Linear Elu (Katharopoulos et al., 2020), Cosformer (Qin et al., 2022), and our proposed Kerformer. The results of these experiments are presented in Figures 7, 8.

The experimental results demonstrate that our proposed method can achieve a faster convergence rate compared to the other four methods, Vanilla attention, Efficient attention, Linear Elu and Cosformer, in the training and validation of the ViT model on the CIFAR-10 dataset. This result fully demonstrates the effectiveness of our proposed method in reducing the training cost of the model.

Compared to traditional attention mechanisms, our proposed improvement achieves better results with less computational cost, indicating that our method can train better models in less time. Therefore, our proposed method has better efficiency and higher performance, making it an effective attention mechanism improvement scheme.

Kerformer provides a good idea of linear complexity by linearizing attention by the operation of activating the Q and K matrices and reweighting the activated K matrices can effectively maintain linear complexity with guaranteed effective attention. In the experimental results Kerformer did not perform best on all tasks, which may be due to the specific nature of the task or the fact that some tasks require a special model structure resulting in poor performance of Kerformer on that task. Also the characteristics of the dataset, the experimental setup, and the choice of hyperparameters may have affected the experimental results of Kerformer on this task.

7. Conclusion

We propose a new Kerformer method to linearize the attention mechanism by the kernel function method to first process the Q and K matrices non-negatively, then reweight the non-negatively processed K matrices by SE Block to amplify the localization relation of the attention matrix, and finally change the order of operations of the attention matrix by the combination law of matrix operation to convert Transformer's computation of the complex attention mechanism into a linear computation based on the sequence length N. We conducted experiments on text classification, Long-range arena, the computational speed of the model on long sequences, and on image classification, respectively, and the experimental results show that Kerformer performs well on these different tasks. This well demonstrates that the Kerformer model can exhibit good model performance and computational efficiency both on NLP tasks and on image tasks, which can make Kerformer widely applicable to different fields where attention mechanisms exist. Overall, our approach can achieve high model performance with low running cost, which allows the deployment of models with attention mechanisms to some devices with low computational power.

In the future, we hope that our proposed method can be widely applied to the computational process of attention mechanism to reduce the running cost of the model, and we will continue to optimize our method so that it can be widely applied to different downstream tasks.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://www.cs.toronto.edu/~kriz/cifar.html; http://yann.lecun.com/exdb/mnist/; https://www.tensorflow.org/datasets?hl=zh-cn.

Author contributions

YG designed the research project, conducted experiments, analyzed the data, and wrote the paper. YF and YL provided guidance and feedback on the research design, data analysis, and paper writing. DW helped with data collection, experiment design, and manuscript proofreading. All authors have read and approved the final manuscript.

Funding

This research was supported by the Research and Development Project of Key Technologies of Edge Computing in the Digital Monitoring Scene (Project) No.: 2022B01008-6.

Acknowledgments

The authors would like to express their gratitude to all individuals and organizations who have provided support and assistance in this research project. We would like to thank our colleagues for their valuable suggestions and feedback. We also thank the reviewers for their insightful comments and suggestions that helped improve the quality of this work. Lastly, the authors also acknowledge the support of the funding agencies that provided financial support for this research.

Conflict of interest

DW was employed by Xinjiang Lianhai INA-INT Information Technology Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: a framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst.* 33, 12449–12460. doi: 10.48550/arXiv.2006.11477

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: the long-document transformer. *arXiv*. [preprint]. doi: 10.48550/arXiv.2004.05150

Bhandare, A., Sripathi, V., Karkada, D., Menon, V., Choi, S., Datta, K., et al. (2019). Efficient 8-bit quantization of transformer neural machine language translation model. *arXiv preprint arXiv:1906.00532*

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., et al. (2020). "End-to-end object detection with transformers," in *Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I 16* (Cham: Springer), 213–229. doi: 10.1007/978-3-030-58452-8_13

Chen, Y., Zeng, Q., Ji, H., and Yang, Y. (2021). Skyformer: remodel self-attention with Gaussian kernel and Nyström method. *Adv. Neural Inf. Process. Syst.* 34, 2122–2135. doi: 10.48550/arXiv.2111.00035

Child, R., Gray, S., Radford, A., and Sutskever, I. (2019). Generating long sequences with sparse transformers. *arXiv*. [preprint]. doi: 10.48550/arXiv.1904.10509

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., et al. (2020). Rethinking attention with performers. *arXiv*. [preprint]. doi: 10.48550/arXiv.2009.14794

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv*. [preprint]. doi: 10.48550/arXiv.1810.04805

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: transformers for image recognition at scale. *arXiv*. [preprint]. doi: 10.48550/arXiv.2010.11929

Kakuba, S., Poulose, A., and Han, D. S. (2022a). Attention-based multi-learning approach for speech emotion recognition with dilated convolution. *IEEE Access* 10, 122302–122313. doi: 10.1109/ACCESS.2022.3223705

Kakuba, S., Poulose, A., and Han, D. S. (2022b). Deep learning-based speech emotion recognition using multi-level fusion of concurrent features. *IEEE Access* 10, 125538–125551. doi: 10.1109/ACCESS.2022.3225684

Katharopoulos, A., Vyas, A., Pappas, N., and Fleuret, F. (2020). "Transformers are RNNs: fast autoregressive transformers with linear attention," in *International Conference on Machine Learning*. (PMLR), 5156–5165.

Kitaev, N., Kaiser, Ł., and Levskaya, A. (2020). Reformer: the efficient transformer. *arXiv*. [preprint]. doi: 10.48550/arXiv.2001.04451

Linsley, D., Kim, J., Veerabadran, V., Windolf, C., and Serre, T. (2018). Learning long-range spatial dependencies with horizontal gated recurrent units. *Adv. Neural Inf. Process. Syst.* 31. doi: 10.32470/CCN.2018.1116-0

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). RoBERTa: a robustly optimized BERT pretraining approach. *arXiv*. [preprint]. doi: 10.48550/arXiv.1907.11692

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., Potts, C., et al. (2011). "Learning word vectors for sentiment analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142–150.

Nangia, N., and Bowman, S. R. (2018). ListOps: a diagnostic dataset for latent tree learning. *arXiv*. [preprint]. doi: 10.48550/arXiv.1804.06028

Ni, J., Li, J., and McAuley, J. (2019). "Justifying recommendations using distantly-labeled reviews and fine-grained aspects," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 188–197. doi: 10.18653/v1/D19-1018

Peng, H., Pappas, N., Yogatama, D., Schwartz, R., Smith, N. A., Kong, L., et al. (2021). Random feature attention. *arXiv*. [preprint]. doi: 10.48550/arXiv.2103.02143

Qin, Z., Sun, W., Deng, H., Li, D., Wei, Y., Lv, B., et al. (2022). cosFormer: rethinking softmax in attention. *arXiv*. [preprint]. doi: 10.48550/arXiv.2202.08791

Radev, D. R., Muthukrishnan, P., Qazvinian, V., and Abu-Jbara, A. (2013). The ACL anthology network corpus. *Lang. Resour. Eval.* 47, 919–944. doi: 10.1007/s10579-012-9211-2

Rymarczyk, D., Borowa, A., Tabor, J., and Zielinski, B. (2021). "Kernel self-attention for weakly-supervised image classification using deep multiple instance learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1721–1730. doi: 10.1109/WACV48630.2021.00176

Scribano, C., Franchini, G., Prato, M., and Bertogna, M. (2023). Dct-former: efficient self-attention with discrete cosine transform. *J. Sci. Comput.* 94, 67. doi: 10.1007/s10915-023-02125-5

Shen, Z., Zhang, M., Zhao, H., Yi, S., and Li, H. (2021). "Efficient attention: attention with linear complexities," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3531–3539.

Tay, Y., Dehghani, M., Abnar, S., Shen, Y., Bahri, D., Pham, P., et al. (2020). Long range arena: a benchmark for efficient transformers. *arXiv*. [preprint]. doi: 10.48550/arXiv.2011.04006

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. Neural Inf. Process. Syst*. 30.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S. R., et al. (2018). GLUE: a multi-task benchmark and analysis platform for natural language understanding. *arXiv*. [preprint]. doi: 10.48550/arXiv.1804.07461

Wang, S., Li, B. Z., Khabsa, M., Fang, H., and Ma, H. (2020). Linformer: self-attention with linear complexity. *arXiv*. [preprint]. doi: 10.48550/arXiv.2006.04768

Wu, Y., Lian, D., Gong, N. Z., Yin, L., Yin, M., Zhou, J., et al. (2021). "Linear-time self attention with codeword histogram for efficient recommendation," *Proceedings of the Web Conference 2021*, 1262–1273. doi: 10.1145/3442381.3449946

Xiong, Y., Zeng, Z., Chakraborty, R., Tan, M., Fung, G., Li, Y., et al. (2021). Nyströmformer: a nyström-based algorithm for approximating self-attention. *Proc. AAAI Conf. Artif. Intell.* 35, 14138–14148. doi: 10.1609/aaai.v35i16.17664

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., et al. (2020). Big bird: transformers for longer sequences. *Adv. Neural Inf. Process. Syst*. 33, 17283–17297. doi: 10.48550/arXiv.2007.14062

Zeng, Z., Xiong, Y., Ravi, S., Acharya, S., Fung, G. M., Singh, V., et al. (2021). "You only sample (almost) once: linear cost self-attention via Bernoulli sampling," in *International Conference on Machine Learning*. PMLR, 12321–12332.

TYPE Original Research
PUBLISHED 31 August 2023
DOI 10.3389/fnbot.2023.1204418

OPEN ACCESS

EDITED BY
Long Jin,
Lanzhou University, China

REVIEWED BY
Dechao Chen,
Hangzhou Dianzi University, China
Rohit Shukla,
University of Wisconsin-Madison, United States

*CORRESPONDENCE
Lang Gao
✉ 1171157437@qq.com

RECEIVED 12 April 2023
ACCEPTED 27 July 2023
PUBLISHED 31 August 2023

CITATION
Ye X, Gao L, Chen J and Lei M (2023) Based on cross-scale fusion attention mechanism network for semantic segmentation for street scenes. *Front. Neurorobot.* 17:1204418.
doi: 10.3389/fnbot.2023.1204418

Based on cross-scale fusion attention mechanism network for semantic segmentation for street scenes

Xin Ye[1], Lang Gao[1]*, Jichen Chen[2] and Mingyue Lei[1]

[1]Institute of Artificial Intelligence and Data Science, Xi'an Technological University, Xi'an, China, [2]Computer Part III, Xi'an Microelectronics Technology Institute, Xi'an, China

Semantic segmentation, which is a fundamental task in computer vision. Every pixel will have a specific semantic class assigned to it through semantic segmentation methods. Embedded systems and mobile devices are difficult to deploy high-accuracy segmentation algorithms. Despite the rapid development of semantic segmentation, the balance between speed and accuracy must be improved. As a solution to the above problems, we created a cross-scale fusion attention mechanism network called CFANet, which fuses feature maps from different scales. We first design a novel efficient residual module (ERM), which applies both dilation convolution and factorized convolution. Our CFANet is mainly constructed from ERM. Subsequently, we designed a new multi-branch channel attention mechanism (MCAM) to refine the feature maps at different levels. Experiment results show that CFANet achieved 70.6% mean intersection over union (mIoU) and 67.7% mIoU on Cityscapes and CamVid datasets, respectively, with inference speeds of 118 FPS and 105 FPS on NVIDIA RTX2080Ti GPU cards with 0.84M parameters.

KEYWORDS

computer vision, semantic segmentation, channel attention mechanism, residual block, dilation convolution, factorized convolution

Introduction

Semantic segmentation is a computer vision task that involves assigning a label to every pixel for a given image based on its content. In the context of street scenes, this task involves identifying and labeling various objects such as buildings, roads, vehicles, and pedestrians.

In the last 10 years, scene understanding has advanced quickly in the fields of computer vision and photogrammetry, particularly the essential task of semantic segmentation (Yang et al., 2021). Semantic segmentation aims to assign a label for each pixel of the images. It has a wide range of applications, including scene comprehension, autonomous vehicle and driver assistance, and augmented reality (Lu et al., 2019). Enabling autonomous cars to be environmentally aware so they can drive safely, and machines to intelligently analyze medical images, reducing the workload for doctors and dramatically reducing the time it takes to run diagnostic tests.

The cross-scale fusion attention mechanism network uses a combination of convolutional neural networks (CNNs) and attention mechanisms to perform semantic segmentation. CNNs are used to extract features from images at multiple scales, while attention mechanisms are used to selectively focus on important regions of the image.

The attention mechanism is an effective way to promote accuracy by computing attention maps that indicate which regions of the feature maps are most relevant for the segmentation task. The attention maps are then used to weigh the features from different scales before they are fused together. This helps to ensure that important information from all scales is taken into account during the segmentation process.

In recent years, deep convolutional neural networks (DCNNs) have demonstrated their amazing capabilities for Image classification tasks. Since the FCN (Long et al., 2015) was proposed, which is the pioneer for semantic segmentation, DCNNs have shown their power in the task of semantic segmentation. It has become the mainstream of segmentation approaches. Compared to traditional visual algorithms, DCNNs achieve good results with their end-to-end approach.

Of course, the development of image segmentation technology also has many shortcomings that need to be improved. With the development trend of artificial intelligence, the network model is getting deeper and bigger. As the network deepens, training will become more and more difficult, mainly because of the gradient explosion in the network training process of gradient descent. Some methods have also been used to improve the situation, such as changing weights and normalization. However, with the deepening of the network model, the training error increases rather than decreases. The emergence of residual networks solves this problem well, and its performance is greatly improved compared to a traditional network.

Most of the prior networks (Long et al., 2015; Badrinarayanan et al., 2017; Chen et al., 2017) neglected the segmentation efficiency while generating outstanding results. They have several disadvantages, including large storage overhead and low computing efficiency. Specifically, they have high computational and storage requirements. Therefore, creating lightweight and efficient networks to solve the above problems is a major trend. The core of our CFANet is ERM with dilated factorized convolution, which can extract features while keeping the computation requirements low. Our main contributions can be summarized as follows:

a) An ERM, which consists of convolutional decomposition and channel shuffling operations, is designed to extract semantic information while keeping the computational cost low.

b) MCAM is introduced to refine the feature maps at different levels.

c) We achieve 70.6% mIoU and 67.7% mIoU on the Cityscapes and CamVid datasets, respectively, along with the inference speed of 118 FPS and 105 FPS on an NVIDIA RTX2080Ti GPU card.

Overall, the cross-scale fusion attention mechanism network is an effective approach the semantic segmentation of street scenes. It has been shown to achieve state-of-the-art performance on several benchmark datasets, demonstrating its potential for real-world applications such as autonomous driving and urban planning.

Materials and methods

In this section, the work related to dilated convolution, factorized convolution and real-time semantic segmentation will be discussed. The following is a general overview of the materials and methods used in the cross-scale fusion attention mechanism network for the semantic segmentation of street scenes:

a) Data Collection: A large dataset of street scenes was collected for training and validation of the neural network. This dataset typically includes high-resolution images and corresponding segmentation masks that label each pixel of the image with the corresponding object or class.

b) Pre-processing: The collected data is pre-processed to prepare it for use in the neural network. This may include resizing the images, normalizing the pixel values, and augmenting the data through techniques such as rotation, flipping, and cropping to increase the size and diversity of the dataset.

c) Network Architecture: The cross-scale fusion attention mechanism network architecture is designed and implemented based on the specific requirements of the semantic segmentation task.

d) Training: The network is trained using the pre-processed data through a process of backpropagation, where the weights of the network are adjusted to minimize the loss function. The training process involves multiple iterations or epochs, where the network is trained on batches of images and corresponding segmentation masks.

e) Evaluation: The performance of the network is evaluated on a separate validation dataset to assess its accuracy and generalization ability. Metrics such as mIoU and pixel accuracy are commonly used to evaluate the performance of the network.

f) Testing: The final step involves using the trained network to perform semantic segmentation on new images in real-world applications. This typically involves feeding the input image through the network and generating a segmentation mask that labels each pixel with the corresponding object or class.

Overall, the materials and methods used in the cross-scale fusion attention mechanism network for semantic segmentation of street scenes involve collecting and pre-processing data, designing and implementing the neural network architecture, training and evaluating the network, and finally testing it in real-world applications.

Dilated convolution

Dilated convolution is a convolutional neural network operation that enables the receptive field of a convolutional layer to be expanded without increasing the number of parameters. It is commonly used in semantic segmentation tasks where the output needs to preserve fine-grained spatial details. In a traditional convolutional layer, each filter kernel slides over the input feature map with a stride of 1, resulting in a receptive field that grows linearly with the kernel size. Dilated convolution, on the other hand, inserts zeros between the kernel values, effectively increasing the kernel's spacing or dilation rate. This means that the receptive

Abbreviations: ERM, Efficient Residual Module; MCAM, Multi-branch Channel Attention Mechanism.

field of the dilated convolutional layer can be increased without increasing the number of parameters.

Dilated convolution is commonly used in deep learning architectures for image analysis, such as in semantic segmentation, where it helps to capture multi-scale features and maintain spatial resolution. It has been shown to improve the performance of neural networks in a variety of computer vision tasks.

For segmented tasks, the feature resolution was decreased due to the consecutive pooling operations or convolution striding. This invariance may have a negative impact on detailed segmentation. To overcome this problem, dilated convolution, which has been proven as an effective way for semantic segmentation tasks. For example, Deeplab (Chen et al., 2017) introduced an atrous spatial pyramid pooling module that applied dilated convolution and pyramid framework to enlarge the receptive field. LedNet (Wang et al., 2019) used dilated convolution in the proposed SS-nbt module to enlarge the efficiency and the accuracy of the residual block. RELAXNet (Liu et al., 2022) applied dilated convolution in the process of the depth separable convolution to compress the module model. All of the above methods demonstrate the effectiveness and lightness of dilated convolution in the segmentation task.

Factorized convolution

In order to improve the inference speed and ensure the segmentation accuracy, factorized convolution is often used to construct lightweight segmentation networks. Factorized convolution is a technique used in deep learning for reducing the computational cost and memory requirements of CNNs. It involves decomposing a standard convolutional operation into two or more separate convolutions, each with a smaller kernel size.

The idea behind factorized convolution is that a large convolutional kernel can be factorized into smaller kernels that are applied sequentially. This reduces the number of parameters in the network and can speed up computation without sacrificing accuracy.

Factorized convolution has several advantages over standard convolutional layers. First, it reduces the number of parameters in the network, which can reduce overfitting and make training faster. Second, it reduces the computational cost of the network by breaking down the convolution into smaller operations. Finally, factorized convolution can improve accuracy in certain cases by allowing for more efficient and targeted feature extraction.

Factorized convolution is commonly used in mobile and embedded deep learning applications where computational and memory resources are limited. It has been shown to be effective in a variety of computer vision tasks, including image classification, object detection, and semantic segmentation.

There are two kinds of factorized methods often used in lightweight networks. One is factorized the standard 3 × 3 convolution into a stacked 1 × 3 and 3 × 1 convolution, and the other is depth separable convolution that factorized the standard convolution into a depth-wise convolution and point-wise convolution. These two factorized methods can dramatically decrease the amount of the parameters. Many real-time semantic segmentation approaches, including FASSD-Net (Rosas-Arias et al., 2021), MDRNet (Dai et al., 2021), and MSCFNet (Gao et al., 2021) use it to construct efficient networks.

Attention mechanisms

Attention mechanisms are a technique used in deep learning to selectively focus on certain parts of the input data during the learning process. It was initially introduced in natural language processing for machine translation, but has since been applied to other domains, including computer vision and speech recognition.

For humans, when we look at a picture, we consciously notice the salient areas and ignore the less important ones. We ask the computer to imitate our behavior, and motivated by this observation, attention mechanisms are introduced into computer vision in order to imitate this aspect of the human visual system. This is the so-called attention mechanism, which is essentially a mechanism for focusing local information. Attention mechanisms have achieved great success in many visual tasks, including image classification, object detection, semantic segmentation, etc.

The idea behind attention mechanisms is to selectively emphasize different parts of the input data, based on their relevance to the task at hand. This is achieved by assigning a weight to each input element, which determines its relative importance. The weights are learned through the training process, allowing the model to adapt to different input patterns. Attention mechanisms are commonly used in neural networks that process sequential or spatial data, such as recurrent neural networks (RNNs) and CNNs. In RNNs, the attention mechanism is typically used to selectively weight different time steps of the input sequence, while in CNNs, it is used to weight different spatial locations in the feature maps. Attention mechanisms have been shown to improve the performance of neural networks in a variety of tasks, including image captioning, machine translation, and speech recognition. It has become a standard component in many state-of-the-art deep learning architectures.

The channel attention mechanism and the spatial attention mechanism are two often used mechanisms. The purpose of using the channel attention module is to make the input image more meaningful. The importance of each channel of the input image is calculated through the network. So as to achieve the purpose of improving the feature representation ability. The attention mechanism (Vaswani et al., 2017) was originally proposed in the natural language field and it assigns each word a different weight. Now, it has been widely used in computer vision tasks. SENet (Hu et al., 2018) generated the feature map weights by modeling the relationship between channels. Besides the channel attention mechanism, CBAM (Woo et al., 2018) used spatial attention mechanisms to assign weights for pixels. The fusion of the high-level and low-level features in the segmentation tasks is an efficient way to improve the accuracy performance. SaNet (Fan and Ling, 2017) introduced a channel shuffle operation for the fusion of the different level features. JPANet (Hu et al., 2022) presented a bilateral path to fuse the feature from different levels.

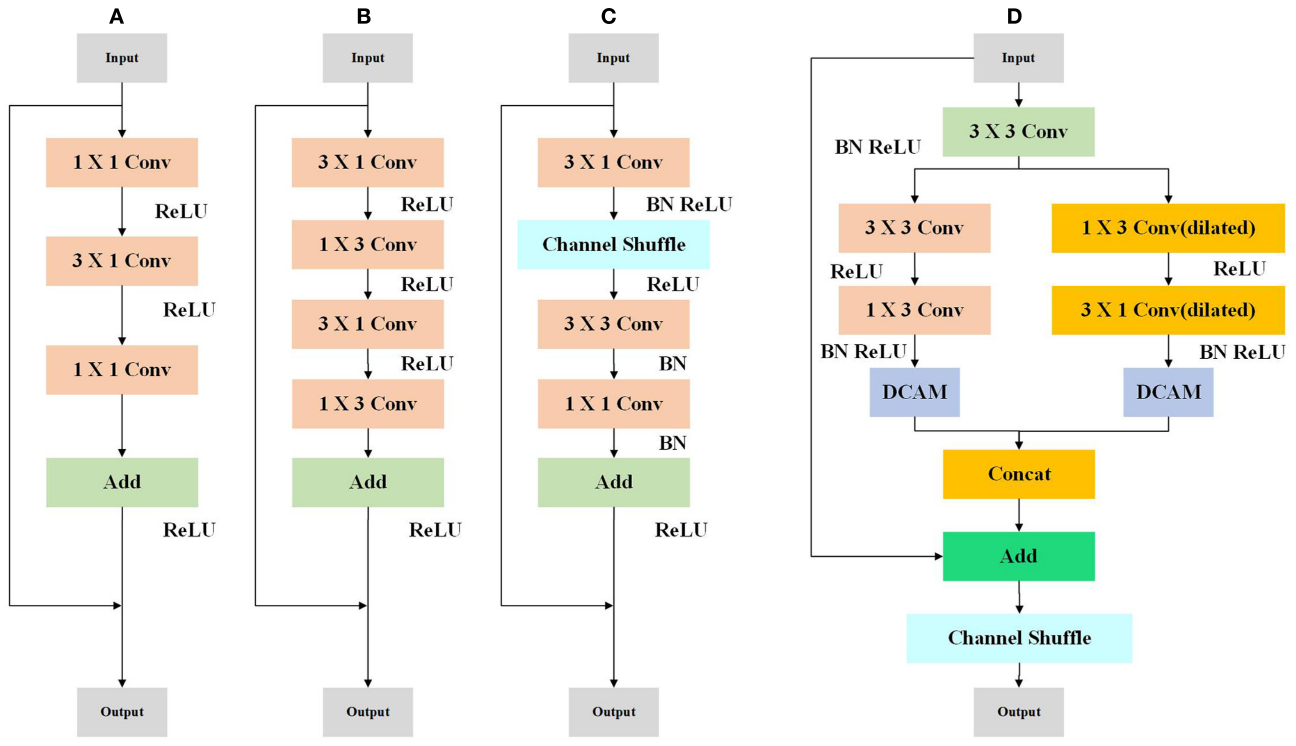


FIGURE 1
Illustration of the bottleneck **(A)**, non-bottle-1d **(B)**, Shufflenet module **(C)**, and ERM **(D)** residual structure.

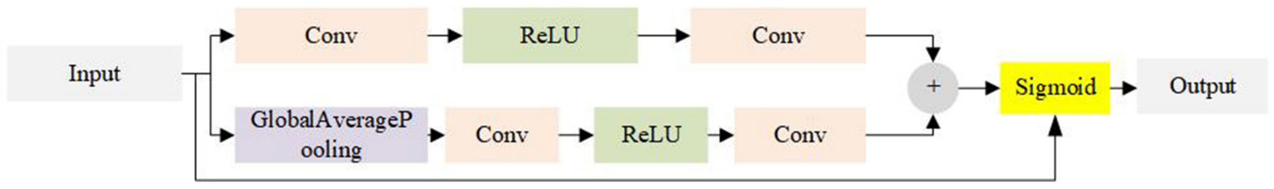


FIGURE 2
The structure of the MCAM.

Methodology

In this section, we first introduce our ERM, which is used for feature extraction.

Subsequently, MCAM is proposed by us. Next, we present the MCAM module that includes the attention mechanism, which is used to fuse features at different levels. At the end of this section, we will discuss the overall architecture of our CFANet, which fuses different levels of features.

Efficient residual module

We concentrate on enhancing the residual structure's effectiveness, which is frequently used in modern CNNs for computer vision tasks. Recent years have seen numerous successful uses of lightweight residual structures, including bottleneck (Figure 1A), non-bottle-1d (Lu et al., 2019) (Figure 1B), and Shufflenet module (Long et al., 2015) (Figure 1C), motivated by LedNet (Wang et al., 2019) and MSCFNet (Gao et al., 2021), We devise an ERM to improve performance with the limitation of computational capacity. Our ERM module is shown in Figure 1D.

In Figure 1, at the beginning of ERM, a standard 3 × 3 convolution is used to decrease the number of the channel by half. The following is a two-branch structure with depth-wise convolution. To be specific, a standard 3 × 3 is divided into consecutive 1 × 3 and 3 × 1 convolutions. The other branch applies dilated depth-wise convolution, which can help enlarge the receptive field. The two-branch is refined by MCAM, which will be introduced in the next subsection.

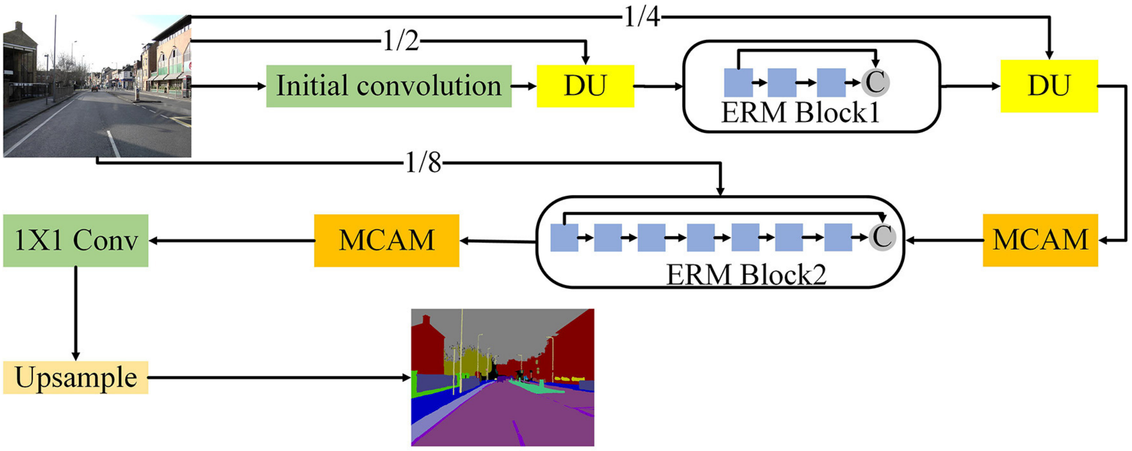


FIGURE 3
Illustration of the overall architecture of the CFANet.

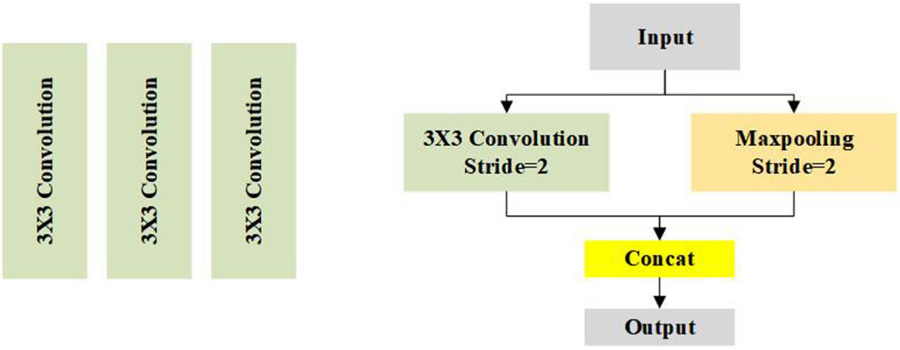


FIGURE 4
Illustration of the initial convolution and down sampling unit.

Multi-branch channel attention mechanism

The attention mechanism can give varying weights to the traits to draw attention to the crucial ones and ignore the unimportant ones. In this paper, we present MCAM to generate different weights for the channels, which is shown in Figure 2.

The convolution is chosen as the local channel context aggregator, which utilizes point-level channel interactions only for each spatial location. As Figure 2 shows, our MCAM module uses global average pooling and 3 × 3 standard convolution in the upper and bottom branches simultaneously. The results from two branches are added element by element. After that, the sigmoid function is used to generate different weights for channels. This procedure can be expressed as follows:

$$MCAM\left(F\right)=F{*}\sigma\left(Add\left(AvgP\left(F\right)+Conv^{3\times3}\left(F\right)\right)\right) \qquad (1)$$

Where F∈ $RC \times H \times W$ denotes the input feature maps, C, H, W represent the channel, height, and width of the feature map, respectively. σ is sigmoid activation function. $\mathrm{Conv}^{3\times3}$ denotes standard convolution with kernel 3×3. Add means the channel wise addition. AvgP is the average pooling operation.

Network architecture design

Based on ERM, we design the architecture of CFANet as shown in Figure 3. In this section, we will introduce the final model of the CFANet.

As can be seen from Figure 3, we first use three 3 × 3 conservative standard convolutions with stride 2 to extract the initial feature of the input images. After the initial convolution, a down sampling unit is used to reduce the size of the feature map and expand the reception domain. However, too many down sampling operations will cause the information, thus, we only employ three down sampling units in our method, thus, the final resolution of the feature map is 1/8 of the input. Our initial convolution and down sampling unit are shown in Figure 4.

The pseudonym code of our CFANet is shown as follows:

```
Input: Image/
Output: The segmentation results
Step 1: Initial Convolution
initial_features = Convolution(input_image,
filters)
Step 2: Fusion and Subsampling
downsampled_image = downsample(input_image,
scale_factor=2)
fusion1 = Concatenate(initial_features,
downsampled_image)
subsampling1 = Subsample(fusion1,
scale_factor=2)
Step 3: Output to ERM Block1
output_ERM_Block1 = ERM_Block1(subsampling1)
Step 4: Fusion and Subsampling
downsampled_image2 =Downsample(input_image,
scale_factor=4)
fusion2 = Concatenate(output_ERM_Block1,
downsampled_image2)
subsampling2 = Subsample(fusion2,
scale_factor=2)
Step 5: MCAM Module
output_MCAM_Module = MCAM_Module(subsampling2)
Step 6: Feature Fusion
fusion3 =Concatenate(output_MCAM_Module,
input_image)
Step 7: MCAM Feature Extraction
output_MCAM_FeatureExtraction =
MCAM_FeatureExtraction(fusion3)
Step 8: 1x1 Convolution
adjusted_features =
Convolution_1x1(output_MCAM_FeatureExtraction,
num_channels)
Step 9: Upsampling
output_feature_map = Upsample(adjusted_features,
scale_factor)
```

Algorithm 1. Cross-scale fusion attention net (CFA-Net).

Experiments

In this part, details and results of our experiments will be presented on the popular semantic segmentation benchmarks Cityscape (Cordts et al., 2016) and CamVid (Brostow et al., 2009). The network was trained on these two data sets, which consisted of high-resolution street view images labeled with pixel-level semantic labels. They used cross-entropy loss functions to train the network and data enhancement techniques such as random scaling and clipping to increase the diversity of the training data. The performance of the proposed network is evaluated using several metrics, including mIoU and pixel accuracy. The results show that the proposed network outperforms several state-of-the-art semantic segmentation networks on the Cityscapes dataset, demonstrating the effectiveness of the cross-scale fusion attention mechanism.

Datasets

Cityscapes dataset

The Cityscapes dataset, contains 19 semantic classes and includes 5,000 fine-labeled samples with the resolution 2,048 × 1,024. The total 5,000 images are divided into training, validation, and test parts. The training parts contain 2,975 images, the validation subset has 500 samples and the test sets have 1,525 images. The sample image and corresponding labels can be seen in Figure 5.

CamVid dataset

The CamVid dataset is collected from a car video sequence, which contains 11 semantic classes and includes 710 labeled images (367 images for training, 101 images for validation, and 233 images for testing). The sample image can be seen in Figure 6.

Data augmentations

In order to overcome the over fitting issue, data enhancement was performed using a horizontal flip and random scale 126. The random scale contains {0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0} Besides, we also use color jitter to adjust the brightness, control, and saturation of the training images and labels.

Training protocols

We train our network with Stochastic Gradient Descent (Bottou, 2010) (SGD) optimizer on Cityscapes dataset with a batch size of 8 on a single NVIDIA RTX2080Ti Card which has 24 GB GPU memory. The learning rate is adjusted by a polynomial policy in the training process. The polynomial policy is computed by $Ircur = init < uscore > Ir \times \left(1 - \frac{epoch}{total<uscore>epoch}\right)^{power}$. The initial learning rate is 4e-2.

When performing training on the CamVid dataset, Adam (Kingma and Ba, 2014) is used as the optimizer with a batch size of 8 and an initial learning rate of 1e-3. We also use a polynomial policy to adjust the learning rate of the training process.

Ablation studies

In this section, the effectiveness of our proposed MCAM was verified by ablation studies. All the ablation experiments are performed on the CamVid dataset, which

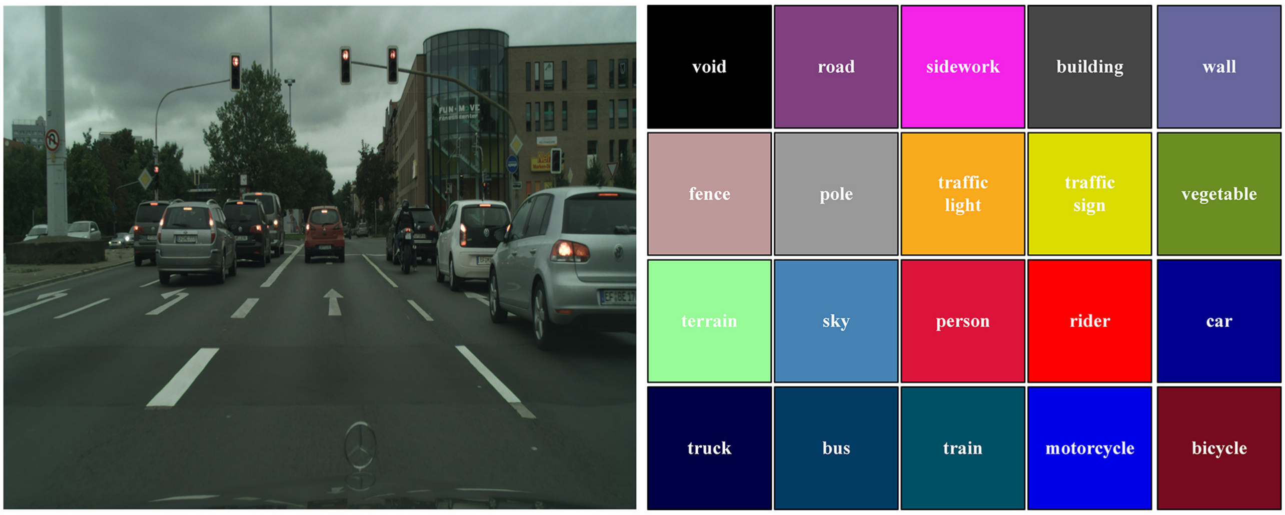


FIGURE 5
The corresponding images and labels of Cityscapes dataset.

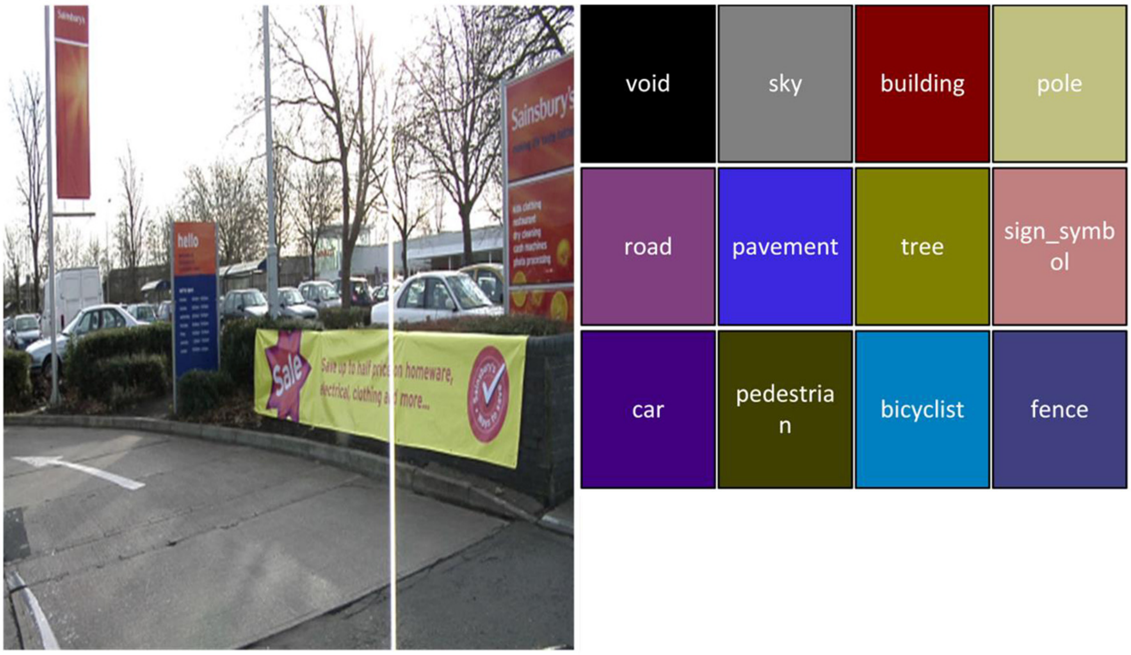


FIGURE 6
The corresponding images and labels of CamVid dataset.

training is time-saving. We trained 1,000 epochs for all the ablation experiments.

Ablation studies on MCAM

In order to prove the effectiveness of MCAM, we removed all the MCAM in our CFANet. The experiment results can be seen in Table 1.

From Table 1, it can be observed that the mIoU decreases by 1% when MCAM is removed. The parameters are reduced to 0.07 million. In other words, our ECAM can effectively increase accuracy with negligible parameters.

TABLE 1 Ablation results on MCAM.

Methods	MCAM	Paramets (*M*)	mIou
CFANet	√	0.84	67.7
CFANet	×	0.77	66.7

TABLE 2 The comprehensive comparisons on Cityscapes dataset.

Method	Input	Backbone	Parameters (*M*)	FLOPs (*G*)	mIoU (%)
SegNet (Badrinarayanan et al., 2017)	640 × 380	VGG16	29.50	286	57.0
Enet (Paszke et al., 2016)	512 × 1,024	No	0.36	3.8	58.3
SQNet (Hu et al., 2018)	1,024 × 2,048	SqueezeNet	–	270	59.8
ESPNet (Mehta et al., 2018)	512 × 1,024	ESPNet	0.36	113	60.3
CGNet (Wu et al., 2020)	360 × 640	No	0.5	-	64.8
ContextNet (Han et al., 2020)	1,024 × 2,048	No	0.85	–	66.1
EDANet (Yang and Gao, 2019)	512 × 1,024	No	0.68	81	67.3
ERFNet (Romera et al., 2017)	512 × 1,024	No	2.10	–	68.0
Fast-SCNN (Zhang et al., 2018)	1,024 × 2,048	No	1.11	–	68.0
BiseNet (Yu et al., 2018)	768 × 1,536	Xception39	5.80	14.8	68.4
ICNet (Zhao et al., 2017)	2,048 × 1,024	PSPNet	26.50	28.3	69.5
DFANet (Li et al., 2019a)	1,024 × 1,024	Xception	7.80	3.4	71.3
Ours	1,024 × 512	No	0.84	10.4	70.6

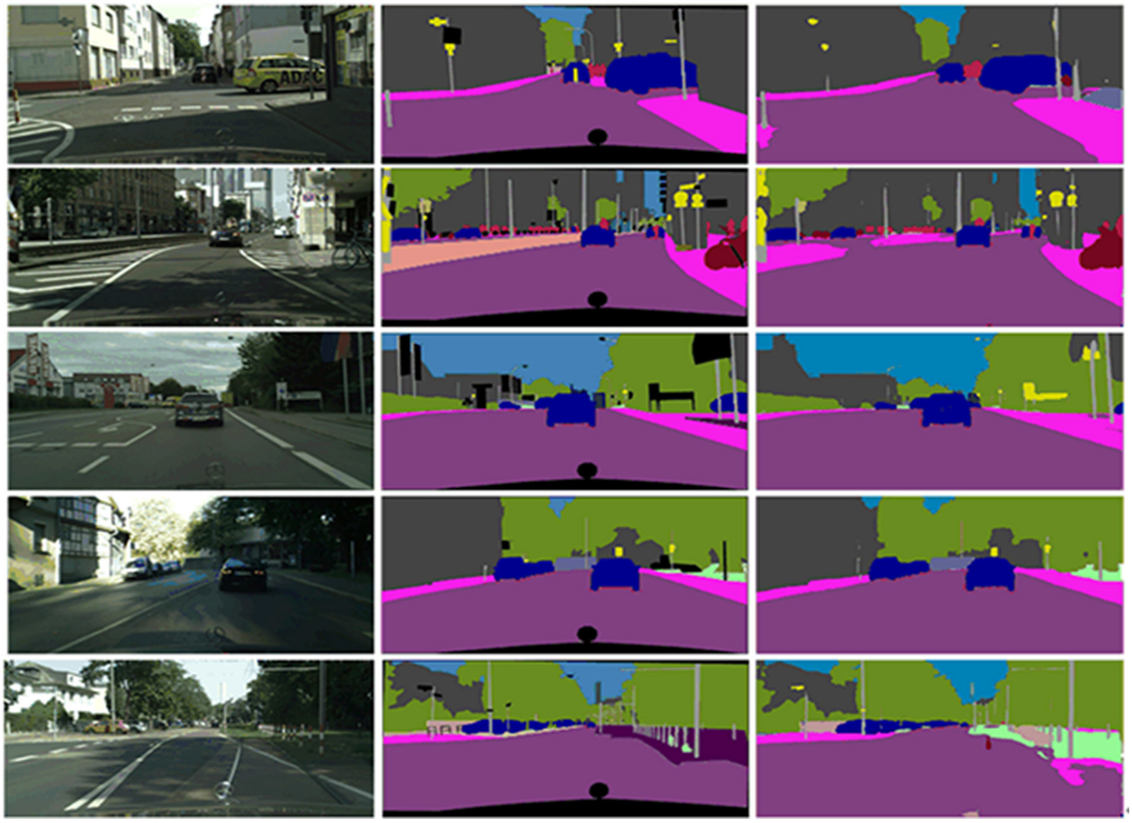


FIGURE 7
The visual results on Cityscapes validation set (from the most-left to right-most is: input, DFANet, and ours).

Performance

In this subsection, Compare our algorithm with the state-of-the-art model. We first report the comparison results on Cityscapes and Camvid benchmarks, then analyze the speed of our model and compute the FPS of other state-of-the-art methods under the same status for fair comparison.

Performance on Cityscapes datasets

A quantitative and quantitative comparison of the urban landscape with other methods is shown. The comparison metrics consist of input size, backbone network, parameter amount, Flops, and the mIoU, the results can be seen in Table2.

It can be observed from Table 2, that the mIoU is comparable to the current state-of-the-art methods, but our CFANet is more lightweight and efficient. The results on Cityscapes show that our approach achieves 71.5% mIoU with only 0.84 million parameters. Compared to DFANet, our method has a similar accuracy but our method only has 0.84 M parameters. Compared to DFANet, our method has a similar accuracy but our method only has 0.84 M parameters. In addition, in order to visualize the results of different methods in terms of segmentation effects, we provide visual comparisons on the Cityscapes validation set. The visual comparison results can be seen from Figure 7.

We also provide a per-class IoU on Cityscapes datasets. Per-class IoU can be seen in Table 3.

Performance on camvid

To further verify the effectiveness of our CFANet, we also evaluated our CFANet on the CamVid dataset. As shown in Table 4, our CFANet obtained remarkable performance against other methods.

From a comprehensive, we select some methods and compared them from four perspectives: input size, backbone, parameter, and mIoU(on test set). As Table 4 shows, our CFANet achieves the best mIoU without backbone. Compared to BiseNet and ICNet, our CFANet is 0.6% higher than ICNet. However, it should be noticed that ICNet has a huge parameter. We provide the visual comparison results of these methods on the CamVid test dataset in Figure 8.

We make a series of supplementary experiments to assess the time performance on an NVIDIA Jeston TX2 platform. The experiment results are shown in Table 5.

A clear comparison is made with other popular algorithms in terms of FLPOS and memory. The results are shown in Table 6.

As shown in Table 6, the memory cost of our CFANet is similar to the ERFNet, but the accuracy performance of our CFANet (in terms of mIoU) is 2.6% higher than it. When compared to the EDANet, the FLOPs of our method are slightly higher than it, but we achieved a 3.3% accuracy promotion, which is significant progress. All the mentioned discussion can prove the effectiveness of our proposed CFANet.

TABLE 3 Per-class IoU(%) performance on the Cityscapes testing set.

Methods	Roa	Sky	Car	Veg	Bui	Sid	Ped	Bus	Tsi	Bic	Ter	TLi	Rid	Pol	Tra	Mot	Wal	Fen	Tru	Ter	mIoU
SegNet (Badrinarayanan et al., 2017)	96.4	91.8	89.3	87.0	84.0	73.2	62.8	43.1	45.1	51.9	63.8	39.8	42.8	35.7	44.1	35.8	28.4	29.0	38.1	63.8	57.0
Enet (Paszke et al., 2016)	96.3	90.6	90.6	88.6	75.	74.0	65.5	50.5	44.0	55.4	61.4	34.1	38.4	43.4	48.1	38.8	32.2	33.2	36.9	61.4	58 3
ESPNet (Mehta et al., 2018)	97.0	92.6	92.3	90.8	76.2	77.5	67.0	52.5	46.3	57.2	63.2	35.6	40.9	45.0	50.1	41.8	35.0	36.1	38.1	63.2	60.3
CGNet (Wu et al., 2020)	95.5	92.9	90.2	89.6	88.1	78.7	74.9	59.5	63.9	60.2	67.6	59.8	54.9	54.1	25.2	47.3	40.0	43.0	44.1	67.6	64.8
ESPNet-v2 (Mehta et al., 2019)	53.0	42.1	43.5	44.2	53.2	49.3	53.1	52.6	66.8	59.9	60.0	65.9	72.9	78.6	88.8	90.5	91.8	93.3	97.3	60.0	66.2
EDANet (Yang and Gao, 2019)	40.9	46.0	42.0	50.4	56.0	52.3	54.3	59.8	68.7	64.0	65.0	58.7	75.7	80.6	89.5	91.4	92.8	94.2	97.7	65.0	67.3
ERFNet (Romera et al., 2017)	97 7	94 2	92 8	91 4	89 8	81 0	76 8	60 1	65 3	61 7	68 2	59 8	57 1	56 3	51 8	47 3	42 5	48 0	50 8	68 2	68 0
ICNet (Zhao et al., 2017)	97.1	93.5	92.6	91.5	89.7	79.2	74.6	72.7	63.4	70.5	8.3	60.4	56.1	61.5	51.3	53.6	43.2	48.9	51.3	8.3	69.5
DABNet (Li et al., 2019b)	97.9	92.8	93.7	91.8	90.6	82.0	78.1	63.7	67.7	66.8	70.1	63.5	57.8	59.3	56.0	51.3	45.5	50.1	52.8	70.1	70.1
LEDNet (Wang et al., 2019)	98.1	94.9	90.9	92.6	91.6	79.5	76.2	64.0	72.8	71.6	61.2	61.3	53.7	62.8	52.7	44.4	47.7	49.9	64.4	61.2	70.6
EdgeNet (Dourado et al., 2020)	98.1	94.9	94.3	92.4	91.6	83.1	80.4	60.9	71.4	67.7	69.7	67.2	61.1	62.6	52.5	55.3	45.4	50.6	50.0	69.7	71.0
Ours	97.2	81.8	90.1	54.1	55.3	59.8	61.8	71.7	92.2	62.2	92.9	77.5	54.2	92.8	52.3	67.1	55.1	51.2	72.1	92.9	70.6

TABLE 4 Comparisons with some of state-of-art methods on CamVid test set.

Methods	Input size	Backbone	Parameter	mIoU
ENet (Paszke et al., 2016)	360 × 480	No	0.36 M	51.3
SegNet (Badrinarayanan et al., 2017)	360 × 480	VGG16	29.5	55.6
NDNet (Yang et al., 2020)	360 × 480	No	0.5	57.2
DFANet (Li et al., 2019a)	720 × 960	Xception	7.8	64.7
Dilation (Rosas-Arias et al., 2021)	720 × 960	VGG16	140.8	65.3
CGNet (Wu et al., 2020)	360 × 480	No	0.5	65.6
BiseNet (Yu et al., 2018)	720 × 960	Xception39	5.8	65.6
DABNet (Li et al., 2019b)	360 × 480	No	0.76	66.4
FDDWNet (Liu et al., 2019)	360 × 480	No	0.80	66.9
ICNet (Zhao et al., 2017)	720 × 960	PSPNet50	26.5	67.1
Ours(CFANet)	360 × 480	No	0.84	67.7

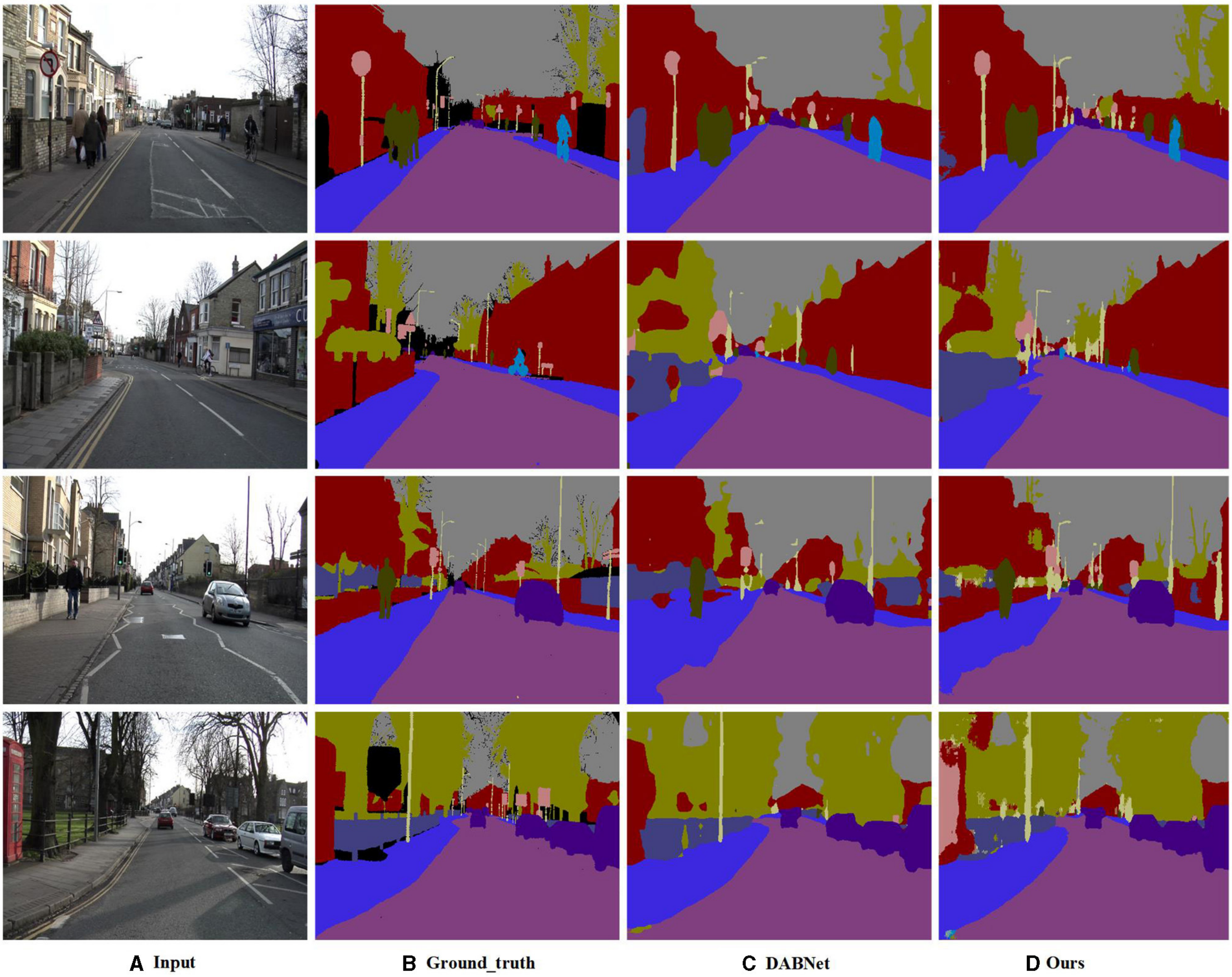


FIGURE 8
The visual results on Camvid testing set. From the most-left to right-most is: Input **(A)**, Ground-Truth **(B)**, DABNet **(C)**, and ours **(D)**.

TABLE 5 The time performance on NVIDIA Jeston TX2.

Method	Input	Platform	FPS	Accuracy mIoU (%)
SegNet (Badrinarayanan et al., 2017)	640 × 480	TX2	5	58
Enet (Paszke et al., 2016)	640 × 480	TX2	26	58.3
EDANe (Yang and Gao, 2019)	640 × 480	TX2	42	67.3
ERFNet (Romera et al., 2017)	640 × 480	TX2	39	68.0
Fast-SCNN (Zhang et al., 2018)	640 × 480	TX2	57	68.0
Ours(CFANet)	640 × 480	TX2	55	70.6

TABLE 6 The comparison results in terms of FLOPS and amount of memory.

Method	Input	Amount of the memory (MB)	FLOPs (G)	Accuracy mIoU (%)
SegNet (Badrinarayanan et al., 2017)	512 × 1,024	1,830	326.26	58
Enet (Paszke et al., 2016)	512 × 1,024	0.36	3.8	58.3
SQNet (Hu et al., 2018)	512 × 1,024	895	270	59.8
ESPNet (Mehta et al., 2018)	512 × 1,024	85	3.2	60.3
CGNet (Wu et al., 2020)	360 × 640	783	6.98	64.8
ContextNet (Han et al., 2020)	512 × 1,024	356	1.78	66.1
EDANet (Yang and Gao, 2019)	512 × 1,024	353	8.95	67.3
ERFNet (Romera et al., 2017)	512 × 1,024	806	25.8	68.0
Fast-SCNN (Zhang et al., 2018)	512 × 1,024	309	1.76	68.0
Ours (CFANet)	512 × 1,024	821	10.4	70.6

Conclusions

In this paper, A new semantic segmentation method, CFANet, is proposed. Which fuses 1/2, 1/4, 1/8 feature maps of the input images. Subsequently, we present a novel ERM consisting of convolution decomposition and dilated convolution. We build our core architecture by using ERM. Besides, we devise MCAM to refine the feature map from different stages. Experiment results show that our method achieves 70.6 and 67.7% mIoU along with 118 FPS and 108 FPS on a single NVIDIA 2080Ti GPU card.

In spite of this, we still have a lot of issues to resolve in the near future. In existing lightweight segmentation models, much useful information is lost in order to obtain the smallest possible model size without compromising accuracy. There is still an unsatisfactory level of segmentation accuracy. Furthermore, the inference speed is not fast enough to process high-resolution images. Additionally, while semantic segmentation networks are extremely important for edge devices, their power consumption is not adequately addressed in existing research. For this reason, we are exploring a novel architecture for semantic segmentation to improve the trade-off between inference speed, accuracy, and power consumption in the future.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

XY: conceptualization, methodology, formal analysis, data curation, project administration, and funding acquisition. LG: software, validation, and visualization. JC: investigation and writing—original draft preparation. ML: resources and supervision. XY and LG: writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the Special scientific research plan of Shaanxi Provincial Department of Education (No. 21JK0684).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 2481–2495. doi: 10.1109/TPAMI.2016.2644615

Bottou, L. (2010). "Large scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010* (Paris: Springer), 177–186.

Brostow, G. J., Fauqueur, J., and Cipolla, R. (2009). Semantic object classes in video: a high-definition ground truth database. *Pattern Recognit. Lett.* 30, 88–97. doi: 10.1016/j.patrec.2008.04.005

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2017). Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 834–848. doi: 10.1109/TPAMI.2017.2699184

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., et al. (2016). "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 3213–3223.

Dai, Y., Wang, J., Li, J., and Li, J. (2021). MDRNet: a lightweight network for real-time semantic segmentation in street scenes. *Assembly Automat.* 46, 725–733. doi: 10.1108/AA-06-2021-0078

Dourado, A., de Campos, T. E., Kim, H., and Hilton, A (2020). "Edgenet: semantic scene completion from rgb-d image," in *2020 25th International Conference on Pattern Recognition (ICPR)* (Milan), 503–510. doi: 10.1109/ICPR48806.2021.9413252

Fan, H., and Ling, H. (2017). "Sanet: structure-aware network for visual trackin," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (Honolulu, HI: IEEE), 42–49.

Gao, G., Xu, G., Yu, Y., Xie, J., Yang, J., Yue, D., et al. (2021). MSCFNet: a lightweight network with multi-scale context fusion for real-time semantic segmentation. *IEEE Transact. Intell. Transport. Syst.* 23, 25489–25499. doi: 10.1109/TITS.2021.3098355

Han, W., Zhang, Z., Zhang, Y., Yu, J., Chiu, C. C., Qin, J., et al. (2020). Contextnet: Improving convolutional neural networks for automatic speech recognition with global context. *arXiv*. 3610–3614. doi: 10.21437/Interspeech.2020-2059

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 7132–7141.

Hu, X., Jing, L., and Sehar, U. (2022). Joint pyramid attention network for real-time semantic segmentation of urban scenes. *Appl. Intell.* 52, 580–594. doi: 10.1007/s10489-021-02446-8

Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv [Preprint].* arXiv: 1412.6980. doi: 10.48550/arXiv.1412.6980

Li, G., Yun, I., Kim, J., and Kim, J. (2019). Dabnet: depth-wise asymmetric bottleneck for real-time semantic segmentation. *arXiv [Preprint].* arXiv: 1907.11357. doi: 10.48550/arXiv.1907.11357

Li, H., Xiong, P., Fan, H., and Sun, J. (2019a). "Dfanet: deep feature aggregation for real-time semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE), 9522–9531.

Liu, J., Xu, X., Shi, Y., Deng, C., and Shi, M. (2022). RELAXNet: residual efficient learning and attention expected fusion network for real-time semantic segmentation. *Neurocomputing* 474, 115–127. doi: 10.1016/j.neucom.2021.12.003

Liu, J., Zhou, Q., Qiang, Y., Kang, B., Wu, X., Zheng, B., et al. (2019). "FDDWNet: a lightweight convolutional neural network for real-time semantic segmentation," in *Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Barcelona: IEEE), 2373–2377.

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA: IEEE), 3431–3440.

Lu, H., Liu, Q., Tian, D., Li, Y., Kim, H., Serikawa, S., et al. (2019). The cognitive internet of vehicles for autonomous driving. *IEEE Netw.* 33, 65–73. doi: 10.1109/MNET.2019.1800339

Mehta, S., Rastegari, M., Caspi, A., Shapiro, L., and Hajishirzi, H. (2018). "Espnet: efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich), 552–568.

Mehta, S., Rastegari, M., Shapiro, L., and Hajishirzi, H. (2019). "Espnetv2: a light-weight, power efficient, and general purpose convolu-tional neural network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9190–9200.

Paszke, A., Chaurasia, A., Kim, S., and Culurciello, E. (2016). Enet: a deep neural network architecture for real-time semantic segmentation. *arXiv [Preprint].* arXiv: 1606.02147. doi: 10.48550/arXiv.1606.02147

Romera, E., Alvarez, J. M., Bergasa, L. M., and Arroyo, R. (2017). Erfnet: efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transact. Intell. Transport. Syst.* 19, 263–272. doi: 10.1109/TITS.2017.2750080

Rosas-Arias, L., Benitez-Garcia, G., Portillo-Portillo, J., Olivares-Mercado, J., Sanchez-Perez, G., Yanai, K., et al. (2021). FASSD-Net: fast and accurate real-time semantic segmentation for embedded systems. *IEEE Transact. Intell. Transport. Syst.* 23, 14339–14360. doi: 10.1109/ICPR48806.2021.9413176

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30, 6000–6010. doi: 10.5555/3295222.3295349

Wang, Y., Zhou, Q., Liu, J., Xiong, J., Gao, G., Wu, X., et al. (2019). "Lednet: a lightweight encoder-decoder network for real-timesemantic segmentation," in *Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP)* (Taipei: IEEE), 1860–1864.

Woo, S., Park, J., Lee, J. Y., and Kweon, I. S. (2018). "Cbam: convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich), 3–19.

Wu, T., Tang, S., Zhang, R., Cao, J., and Zhang, Y. (2020). Cgnet: a light-weight context guided network for semantic segmentation. *IEEE Transact. Image Process.* 30, 1169–1179. doi: 10.1109/TIP.2020.3042065

Yang, C., and Gao, F. (2019). "EDA-Net: dense aggregation of deep and shallow information achieves quantitative photoacoustic blood oxygenation imaging deep in human breast," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer), 246–254.

Yang, M. Y., Kumaar, S., Lyu, Y., and Nex, F. (2021). Real-time semantic segmentation with context aggregation network. *ISPRS J. Photogr. Remote Sens.* 178, 124–134. doi: 10.1016/j.isprsjprs.2021.06.006

Yang, Z., Yu, H., Fu, Q., Sun, W., Jia, W., Sun, M., et al. (2020). NDNet: Narrow while deep network for real-time semantic segmentation. *IEEE Transact. Intell. Transport. Syst.* 22, 5508–5519. doi: 10.1109/TITS.2020.2987816

Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N., et al. (2018). "Bisenet: bilateral segmentation network for real-time semantic seg-mentation," in *Proceedings of the European Conference on Computer Vision* (Munich), 325–341.

Zhang, X., Chen, Z., Wu, Q. J., Cai, L., Lu, D., Li, X., et al. (2018). Fast semantic segmentation for scene perception. *IEEE Transact. Ind. Informat.* 15, 1183–1192. doi: 10.1109/TII.2018.2849348

Zhao, H., Qi, X., Shen, X., Shi, J., and Jia, J. (2017). "Icnet for real-time semantic segmentation on high-resolution images," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Venice), 405–420.

frontiers | Frontiers in Neurorobotics

TYPE Original Research
PUBLISHED 20 September 2023
DOI 10.3389/fnbot.2023.1242063

OPEN ACCESS

EDITED BY
Long Jin,
Lanzhou University, China

REVIEWED BY
Robert W. Newcomb,
University of Maryland, United States
Guibing Zhu,
Zhejiang Ocean University, China
Pengfei Zhang,
Huzhou University, China

*CORRESPONDENCE
Yuxuan Cao
✉ yuxuancao@foxmail.com

RECEIVED 18 June 2023
ACCEPTED 22 August 2023
PUBLISHED 20 September 2023

CITATION
Cao Y, Liu B and Pu J (2023) Robust control for a tracked mobile robot based on a finite-time convergence zeroing neural network.
Front. Neurorobot. 17:1242063.
doi: 10.3389/fnbot.2023.1242063

Robust control for a tracked mobile robot based on a finite-time convergence zeroing neural network

Yuxuan Cao*, Boyun Liu and Jinyun Pu

College of Power Engineering, Naval University of Engineering, Wuhan, China

Introduction: Since tracked mobile robot is a typical non-linear system, it has been a challenge to achieve the trajectory tracking of tracked mobile robots. A zeroing neural network is employed to control a tracked mobile robot to track the desired trajectory.

Methods: A new fractional exponential activation function is designed in this study, and the implicit derivative dynamic model of the tracked mobile robot is presented, termed finite-time convergence zeroing neural network. The proposed model is analyzed based on the Lyapunov stability theory, and the upper bound of the convergence time is given. In addition, the robustness of the finite-time convergence zeroing neural network model is investigated under different error disturbances.

Results and discussion: Numerical experiments of tracking an eight-shaped trajectory are conducted successfully, validating the proposed model for the trajectory tracking problem of tracked mobile robots. Comparative results validate the effectiveness and superiority of the proposed model for the kinematical resolution of tracked mobile robots even in a disturbance environment.

KEYWORDS

tracked mobile robot, trajectory tracking, finite-time convergence, zeroing neural network, robust

1. Introduction

At present, robots are being widely used in marine exploration (Fang et al., 2022; Wang et al., 2022), industrial manufacturing (Šegota et al., 2021; Truong et al., 2021), military applications (Bistron and Piotrowski, 2021; Rawat et al., 2021), and other fields. Tracked mobile robots (TMRs) show their wide adaptability and traffic ability to complex terrain (Gu et al., 2021). The demand for their motion autonomy and intelligence is increasing. Therefore, the control issue of trajectory tracking has been a research hotspot.

However, a TMR is a typical nonlinear system, and its model parameters change with its motion. In addition, the model is vulnerable to various interferences. The superposition of many factors poses a great challenge to the control algorithm. Therefore, a feasible solution with outstanding convergence performance as well as robustness to handle the nonlinear time-varying control issue of the TMR is imperative in practice. Numerous methodologies and techniques for addressing the tracking control issues of robot systems have been extensively studied and reported, including backstepping control (Ji et al., 2002; Gao et al., 2022; Sabiha et al., 2022), sliding mode control (Ahmed et al., 2021; Yin et al., 2021), fuzzy control (Lara-Molina and Dumur, 2021; Li et al., 2022), and neural network (Ding et al., 2018; Jin and Qiu, 2022).

Among various kinds of solutions, neural network approaches have shown huge advantages in terms of parallelism and easy implementation by hardware (Chen and Zhang, 2018). As a powerful approach for solving time-varying problems, the conventional zeroing neural network (CZNN) proposed in Zhang et al. (2002) has been thoroughly investigated in recent years (Miao et al., 2015; Xiao et al., 2017; Gerontitis et al., 2022; Sun et al., 2022; Zhang and Zheng, 2022). Ma et al. (2021) proposed a new ZNN model to solve the bound-constrained time-varying nonlinear equation, which has been applied to the mobile robot manipulator. Chen et al. proposed a multi-constrained ZNN. The application on the mobile manipulator for nonlinear optimization control demonstrated its physical effectiveness (Chen et al., 2021). Although CZNN can converge to the analytical solution with time, the convergence time is infinite in theory, which is impossible in reality. For an actual situation, the convergence time should be as short as possible. Moreover, CZNN is sensitive to noise and other disturbances. However, the system is susceptible to external disturbances and possible internal disturbances.

Many efforts have been made to address the shortcomings of CZNN. Hu et al. (2020) developed a noise tolerance ZNN model, which successfully tracked the desired path of the mobile manipulator with high accuracy under perturbation. Chen and Zhang (2018) proposed a robust ZNN model for solving the inverse kinematics problem of mobile robot manipulators . Luo et al. proposed a new hyperbolic tangent varying-parameter ZNN. Furthermore, trajectory tracking tasks of the mobile robot substantiate the outstanding convergence of hyperbolic tangent variant-parameter robust ZNN (HTVPR-ZNN) schemes (Luo et al., 2022). Chen et al. (2020) proposed a ZNN model with a super twisting algorithm that realized finite-time convergence and anti-disturbance, proving its effectiveness and superiority in the tracking control of the mobile robot manipulator. Lin et al. utilized a new design formula of noise resistance and finite-time convergence to establish a new ZNN. Compared with CZNN, the presented model was nonsensitive to various types of external disturbances (Xiao et al., 2019). Yan et al. (2019) proposed several improved ZNN models that allow nonconvex activation functions and have accelerated finite-time convergence.

However, the models and approaches reviewed above might potentially not be time-efficient and simultaneously robust for direct applications to a tracking control problem of TMR due to the requirement of timeliness as well as the influence of the disturbance environment. Moreover, it is worth pointing out that the robustness and finite convergence of ZNN models are related to the design of appropriate activation functions. The sign-bi-power function mentioned above endows ZNN with finite-time convergence, but it also contains a sign function, which may lead to singularity and discontinuity. Additionally, the performance under disturbance has not been not fully studied. Therefore, it is necessary to design a new activation function to obtain anti-interference and outstanding convergence.

Under the framework of the ZNN, a finite-time convergence ZNN, termed FCZNN, is proposed in this study. First, a new fractional evolution formula is designed to accelerate the convergence speed and enhance its robustness, which can converge to the desired trajectory within a finite-time under four common disturbances. To better demonstrate the contribution of this study, some existing models are introduced for comparison to highlight the main differences, and the corresponding comparison results are presented in Section 4.

The rest of this paper is organized into four sections. Section 2 presents a novel tracking control method based on FCZNN models for TMR. Section 3 validates the finite convergence and other properties. Section 4 illustrates the corresponding simulation results of the proposed method and presents some existing models for comparison. Section 5 concludes the entire paper.

Before ending this section, the main contributions of this study are summarized as follows:

- A new fractional exponential activation function is proposed in this study and investigated to solve the trajectory tracking issue. Compared with the tunable activation function, the singularity and sign function can be effectively avoided by reasonably selecting the design parameters.
- The finite-time convergence and robustness of the proposed FCZNN are validated theoretically based on the Lyapunov stability theory.
- Simulation experiments are conducted to present the verification and superiority of the FCZNN when compared with some existing models. Additionally, the validity of the theoretical analysis is confirmed based on the corresponding results.

2. Preliminaries

Since the actual situation is complicated, it is difficult to reflect it fully. Appropriate simplification is necessary. First, the main application scenario of our TMR is in a structured environment, such as indoors or on roads, and it can be analyzed on a two-dimensional plane. Furthermore, the difference in grounding pressure and the mass distribution of the TMR affect the kinematic model of the TMR. To simplify the kinematics model, some assumptions are declared for the TMR:

Assumption 1. *The TMR moves on the flat terrain with even tracking grounding pressure.*

Assumption 2. *The centroid of the TMR is located at the center of the robot.*

In the global XOY coordinate system, the schematic diagram of the motion of the TMR is presented in Figure 1. Some notations mentioned in Figure 1 are listed in Table 1.

First, we introduce a model-free tracking control method for the TMR relying only on user-defined input and sensory output without knowing any information about the model parameters of the TMR. The kinematics model of the TMR is depicted as

$$\dot{q}(t) = J(\theta)u(t) \tag{1}$$

where $q(t) = [x, y, \theta]^T$ is the generalized coordinates of the TMR, $\dot{q}(t)$ is the time-derivative of $q(t)$, $u(t) = [v, \omega]^T$ is the

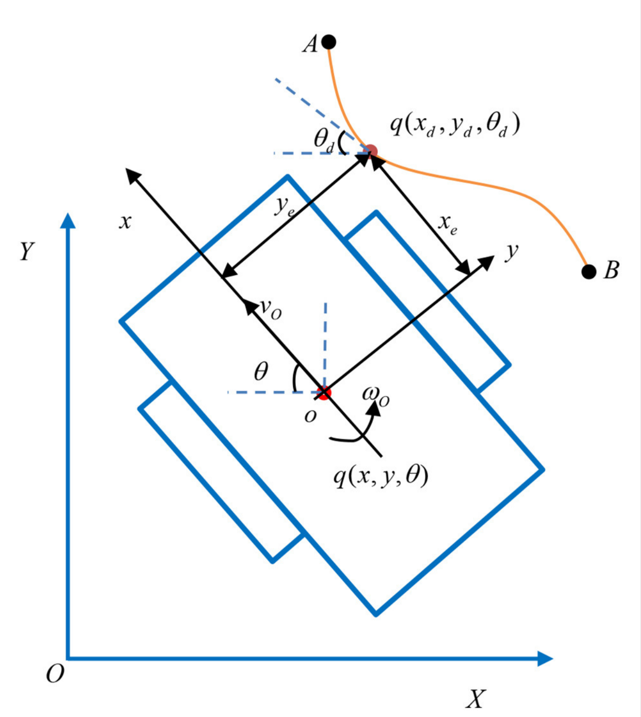


FIGURE 1
Schematic diagram of the motion of the TMR.

TABLE 1 Notations in Figure 1.

Notation	Meaning
xoy	The coordinate system attached to the TMR
$q(x, y, \theta)$	The actual position
$q(x_d, y_d, \theta_d)$	The desired position
o	The centroid of the TMR
θ	The heading angle of the TMR
v_o	The velocity of the TMR
ω_o	The angular velocity of the TMR

control input vector, and $J(\theta) = [\cos\theta, 0; \sin\theta, 0; 0, 1]$ is the full-rank velocity transformation matrix. To obtain the solution of the matrix equation, the FCZNN model is presented to solve this kind of a robot trajectory control issue.

A time-varying desired path equation $q_d(t)$ is offered for tracking using the TMR,

$$\dot{q}_d(t) = J(\theta_d)u_d(t) \tag{2}$$

where $\dot{q}_d(t)$ denotes the time derivate of $q_d(t)$, $J(\theta_d) = [\cos\theta_d, 0; \sin\theta_d, 0; 0, 1]$ is the desired full-rank velocity transformation matrix, and $u_d(t) = [v_d, \omega_d]^T$ is the desired control input vector. The mapping relation in real time t is expressed as $q(t) \rightarrow q_d(t)$. The mapping at the velocity level is shown as $\dot{q}(t) \rightarrow \dot{q}_d(t)$

The following error function is defined in the global coordinate system:

$$q_e(t) = q(t) - q_d(t) \tag{3}$$

The error is generally defined in the coordinate system of then TMR; then, one has

$$e(t) = Tq_e(t) \tag{4}$$

where $e(t) = [e_x, e_y, e_\theta]^T$ and $T = [\cos\theta, \sin\theta, 0; \sin\theta, -\cos\theta, 0; 0, 0, 1]$ is the coordinate transformation matrix, which converts the tracking error defined under the inertial coordinate system to the body coordinate system.

In view of the design rules of the ZNN, the following formula is given:

$$\frac{de(t)}{dt} = -\Gamma\Phi(e(t)) \tag{5}$$

where $\Phi(e(t))$ denotes an activation function vector with various type, linear type, power type, etc. Theoretically, any monotonically increasing odd function can be the activation function candidate. Γ is a positive-definite matrix for scaling the convergence rate of the solve process. Based on the related derivate theory, Γ should be set as large as possible within the tolerance limit of the hardware. For ease of discussion, Γ is set as a diagonal matrix with the same element, that is, $\Gamma = \gamma I$, where I is the identity matrix. Additionally, Γ is a constant scalar-valued parameter matrix. Then,

$$\dot{e}(t) = -\gamma\Phi(e(t)) \tag{6}$$

where γ is the parameter that adjusts the convergence rate.

Moreover from (13), one promtly has

$$\dot{e}(t) = \dot{T}q_e(t) + T\dot{q}_e(t) \tag{7}$$

3. Model design and theoretical analysis

In this section, a finite-time and robust unified framework synthesized by adopting a new activation function is proposed. The relative theorems and proofs about the corresponding features, namely, of finite-time convergence, global stability, and robustness in the disturbance environment, are explored to demonstrate the effectiveness of the proposed FCZNN model.

Considering (1), (6), and (7), one can obtain

$$T(J(\theta)u(t) - \dot{q}_d(t)) + \dot{T}(q(t) - q_d(t)) = -\gamma\Phi(e(t)) \tag{8}$$

Evidently, the neural dynamics Equation (8) makes full use of the pose information and its derivate of the TMR, which contributes to solving the trajectory tracking control problem.

To demonstrate the anti-interference performance of the proposed FCZNN, some theorems about robustness are investigated in this section. Generally, the synthesized error caused by the disturbances is inevitable for any electronic system and neural dynamics. The synthesized error caused by hardware implementation off-set errors can be treated as dynamic non-disappearing noise in linear or sine form. The one caused by the instantaneous decline of power sources or other external disturbances can be regarded as dynamic disappearing noise in exponential form. Then, the implicit dynamic Equation (8) with the synthesized error is reformulated

$$T(J(\theta)u(t)-\dot{q}_d(t))+\dot{T}(q(t)-q_d(t))=-\gamma\Phi(e(t))+W(t) \quad (9)$$

where $W(t) \in R^3$ denotes the synthesized error (could be constant or time-varying) with each entry $w_i(t) \leq w$ for $i = 1, 2, 3$, where $w \geq 0$ is an unknown constant.

3.1. Design of the FCZNN

As mentioned before, the choice of error evolution formula has a crucial influence on the characteristics of the system. Inspired by Xiao et al. (2017), a new fractional exponential activation function is proposed for constructing the error evolution formula.

$$\Phi(x)=\kappa_1 f^{p/p_1}(x,t)+\kappa_2 f(x,t)+\kappa_3 f^{p_1/p}(x,t) \quad (10)$$

where $f(x, t)$ is the set of increasing odd functions and design parameters p and p_1 denote positive odd integer with $p > p_1$, $\kappa_1 > 0, \kappa_2 > 0$, $\kappa_3 > 0$. Evidently, three terms of the activation function are odd functions the sum of the three terms is still a monotonically increasing odd function. For analysis, we define $f(x, t) = x$. Then, the error evolution formula is given as

$$\frac{de(t)}{dt}=-\gamma\left(\kappa_1 e^{p/p_1}(t)+\kappa_2 e(t)+\kappa_3 e^{p_1/p}(t)\right) \quad (11)$$

where γ is defined as before. The Equation (9) can be reformulated as

$$u(t)=J^{\dagger}(\theta)T^{-1}[-\gamma\Phi(e(t))-\dot{T}(q(t)-q_d(t))+T\dot{q}_d(t)+W(t)] \quad (12)$$

where $J^{\dagger}(t)$ denotes the pseudo inverse of $J(t)$.

```
Initialize: TMR initial state vector combined velocity vector $\dot{q}(0)$;
Choose: The tracking duration $T_f$ and design parameters $\gamma$ and $\kappa_{i=1,2,3}$ ;
Input: The desired position $q_d(t)$ of tracking task;
if $t < T_f$ then
    Calculate: The desired path as $\dot{q}_d(t)$ ;
    Read: The real time TMR actual position $q(t)$;
    Calculate: The control-signal by using neuron dynamic equation
        $u(t)=J^{\dagger}(\theta)T^{-1}[-\gamma\Phi(e(t))-\dot{T}(q(t)-q_d(t))+T\dot{q}_d(t)+W(t)]$
    Update: The TMR position in the next moment
    Output: The actual trajectory $q(t)$
else
    Stop: TMR trajectory tracking task finished.
endif
```

Algorithm 1. Tracking control of the TMR via the FCZNN.

The detailed algorithm description about the FCZNN model for the TMR tracking control issue is presented in Algorithm 1. The block diagram presented in Figure 2 demonstrates the principle of the control strategy.

To illustrate the details of the proposed model, the ith ($i = 1, 2, 3$) neuron of the FCZNN is given below.

$$\dot{q}_i=-\gamma\phi(e_i)+w_i-\sum\nolimits_{j=1}^{3}(\dot{T}_{ij}e_j+T_{ij}\dot{q}_{dj}) \quad (13)$$

where $\dot{q}_i$, $\dot{q}_{dj}$ denote the ith element of $\dot{q}$, $\dot{q}_d$, respectively, and $\dot{T}_{ij}$, T_{ij} are the (i, j)th element of $\dot{T}$ and T.

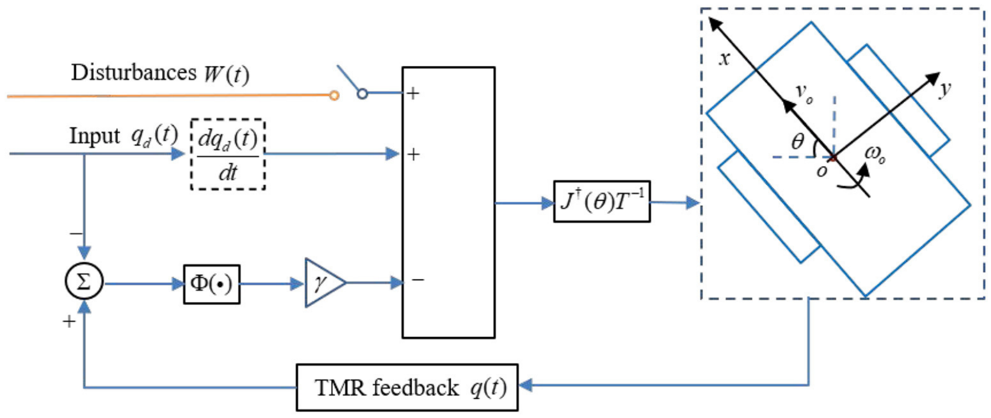


FIGURE 2
Block diagram of the FCZNN model with the possible disturbances for $W(t)$ handling tracking control issue of the TMR.

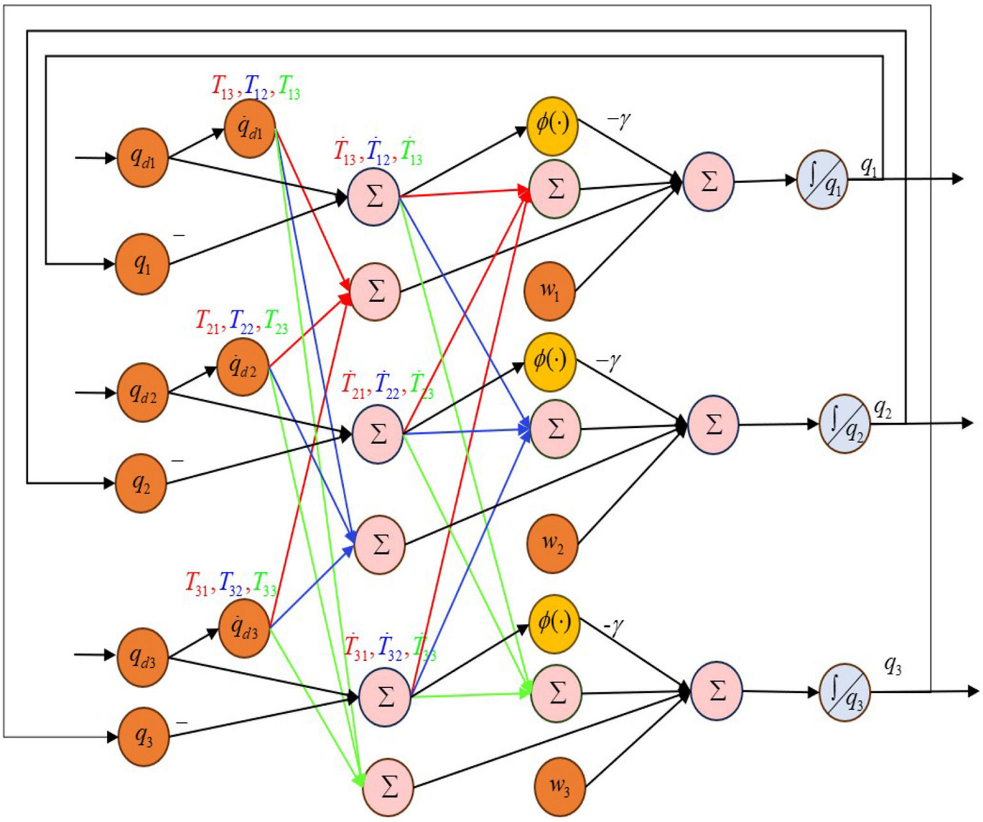


FIGURE 3
Neural topology of the proposed FCZNN model.

Based on (13), the neural topology structure of the proposed FCZNN model is presented in Figure 3.

3.2. Convergence analysis

3.2.1. Global stability analysis

Theorem 1. *If a monotonically increasing odd function $\Phi(\cdot)$ is taken as the activation function, the output will globally converge to the desired trajectory $q_d(t)$ of the model (9) with a random generated initial state $q(0)$.*

Proof: To prove the global convergence of the model (9), the following Lyapunov function candidate is presented as

$$L(t)=\frac{\left\|e(t)\right\|_2^2}{2}=\frac{e^T(t)e(t)}{2} \tag{14}$$

where $\|\cdot\|_2$ denotes the two norm of a vector. Considering (6), the derivate of the above function is

$$\begin{aligned}\dot{L}(t)&=e^T(t)\tfrac{e(t)}{dt}\\&=-\gamma e^T(t)\Phi(e(t))\\&=-\gamma\sum_{i=1}^{m}e_i\phi(e_i(t))\end{aligned} \tag{15}$$

where $e_i(t)$ is the ith element of $e(t)$, $\phi(e_i(t))$ is the ith element of $\Phi(e(t))$, and $m=3$ represents the number of model subsystems. Since the activation function is an odd function, the following relationship exists:

$$e_i(t)\phi(e_i(t))=\begin{cases}>0, & if \;\; e_i(t)\neq 0\\ =0, & if \;\; e_i(t)=0\end{cases}. \tag{16}$$

According to the Lyapunov stability theory, the system is asymptotically stable at moment t with $\dot{L}(t) < 0$ guaranteed. Considering (16), we have

$$\dot{L}(t)=-\gamma\sum_{i=1}^{m}e_i\phi(e_i(t))=\begin{cases}=0 & if \;\; e_i(t)=0\\ <0, & if \;\; e_i(t)\neq 0\end{cases}, \quad t\in[0, \;\; +\infty) \tag{17}$$

□

Equation (17) demonstrates that $\dot{L}(t)$ is negative finite. Based on the Lyapunov stability theory, the system will gradually stabilize with time, the error equation will converge to 0, and the corresponding input will converge to the analytical solution. The proof of global convergence is thus completed.

Theorem 1 indicates that the system residual error converges to 0, which means that the TMR can track the desired position with time. The evolution formula proposed in this study demonstrates that the tracking task of a desired path can converge in the finite time. Next, the finite-time convergence of the FCZNN is proved below.

3.2.2. Finite-time convergence analysis

Theorem 2. *Considering the novel activation function (10) for the error function* $e(t)$, $e(t)$ *can converge to 0 in finite time* T_f. T_f *satisfies the following inequality:*

$$T_f \le \begin{cases} \frac{1}{2\gamma\kappa_2\left(\frac{(p+p_1)}{2p_1}-1\right)} \ln\left(\frac{\kappa_1+\kappa_2}{\kappa_2 L(0)^{1-\frac{p+p_1}{2p_1}}+\kappa_1}\right) + \frac{\ln\left(1+\frac{\kappa_2}{\kappa_3}L(0)^{1-\frac{(p+p_1)}{2p}}\right)}{2\gamma\kappa_2\left(\frac{(p+p_1)}{2p}-1\right)}, L(t) \ge 1 \\ \frac{\ln\left(1+\frac{\kappa_2}{\kappa_3}L(0)^{1-\frac{(p+p_1)}{2p}}\right)}{2\gamma\kappa_2\left(\frac{(p+p_1)}{2p}-1\right)}, L(t) < 1 \end{cases} \tag{18}$$

Proof: Firstly, the maximum initial value element of the error function is depicted as $e^{+}(0) = \max_{i=1,2,3}\left\{\left|e_i(0)\right|\right\}$. The following relationship holds true: $-\left|e^{+}(t)\right| \le \left|e_i(t)\right| \le \left|e^{+}(t)\right|$ for $t \ge 0$ and $i = 1,2,3$, which reveals that $e_i(t)$ converges to 0 when $e^{+}(t)$ is equivalent to 0. Moreover, $\dot{e}^{+}(t) = -\gamma\Phi\left(e^{+}(t)\right)$.

$$\begin{aligned} \dot{L}(t) &= 2\dot{e}^{+}(t)e^{+}(t) \\ &= -2\gamma\Phi\left(e^{+}(t)\right)e^{+}(t) \\ &= -2\gamma\left(\kappa_1 L(t)^{(p+p_1)/2p_1} + \kappa_2 L(t) + \kappa_3 L(t)^{(p_1+p)/2p}\right) \end{aligned} \tag{19}$$

For simplicity, we define $2\gamma\kappa_1 = \beta_1$, $2\gamma\kappa_2 = \beta_2$, $2\gamma\kappa_3 = \beta_3$, $a = (p+p_1)/2p_1$, and $b = (p+p_1)/2p$. In view of the precondition, $a > 1$, $0 < b < 1$. Then, $\dot{L}(t) = -(\beta_1 L^a(t) + \beta_2 L(t) + \beta_3 L^b(t))$.

Inequality (18) is proved below. The following two situations exist:

CASE I: When $L(t) \ge 1$,

$$\dot{L}(t) \le -\beta_1 L^a(t) - \beta_2 L(t) \tag{20}$$

Inequality (20) can be transformed as

$$\frac{dL(t)}{\beta_3 L^a(t) + \beta_2 L(t)} \le -dt \tag{21}$$

Integrating both sides of (21) from 0 to t, we can obtain

$$L(t) = \begin{cases} \le \exp(-\beta_2 t)\left(L^{1-a}(0) + \frac{\beta_1}{\beta_2} - \frac{\beta_1}{\beta_2}\exp((1-a)\beta_2 t)\right)^{\frac{1}{1-a}}, if\ \ 0 \le t < t_1 \\ = 1, if\ \ t = t_1 \end{cases} \tag{22}$$

where t_1 denotes the convergence time to 0 for $L^{1-a}(0)=\max_{i=1,2,3}\left\{e_i^{1-a}(0)\right\}$.

Let $L(t) = 1$,

$$t_1 = \frac{1}{(a-1)\beta_2}\ln\frac{\beta_1+\beta_2}{\beta_2 L^{1-a}(0)+\beta_1}. \tag{23}$$

CASE II: When $L(t) \le 1$,

$$\dot{L}(t) \le -(\beta_2 L(t) + \beta_3 L^b(t)). \tag{24}$$

Inequality (24) can be converted to

$$\frac{dL_2(t)}{\beta_2 L_2(t) + \beta_3 L_2^b(t)} \le -dt. \tag{25}$$

Integrating the above differential inequality from 0 to t, we have

$$L(t) = \begin{cases} \le \exp(-\beta_2 t)\left(L^{1-b}(0) + \frac{\beta_3}{\beta_2} - \frac{\beta_3}{\beta_2}\exp((1-b)\beta_2 t)\right)^{\frac{1}{1-b}}, if\ \ 0 \le t < t_2 \\ = 0, if\ \ t = t_2 \end{cases}. \tag{26}$$

Similarly, t_2 satisfies the following equality:

$$t_2 = \frac{\ln\left(1+\frac{\beta_2}{\beta_3}L^{1-b}(0)\right)}{\beta_2(1-b)}. \tag{27}$$

where t_2 denotes the convergence time to 0 for $L(t) \le 1$, and $L^b(0)=\max_{i=1,2,3}\left\{e_i^b(0)\right\}$.

In summary, the upper bound of convergence time T_f satisfies

$$T_f \le \begin{cases} \frac{1}{(a-1)\beta_2}\ln\left(\frac{\beta_1+\beta_2}{\beta_2 L^{1-a}(0)+\beta_1}\right) + \frac{\ln\left(1+\frac{\beta_2}{\beta_3}L^{1-b}(0)\right)}{\beta_2(1-b)},\ \ L(t) \ge 1 \\ \frac{\ln\left(1+\frac{\beta_2}{\beta_3}L^{1-b}(0)\right)}{\beta_2(1-b)},\ \ L(t) < 1 \end{cases}. \tag{28}$$

Note that (28) can be rewritten in the form of (18). The proof is thus completed. □

3.3. Robustness analysis

The CZNN has been proven to converge to the desired result in the disturbance-free case. However, in the practical situation, the disturbance cannot be avoided. The tracking error may arise in the presence of the disturbance. In this section, the steady-state error is given base on the Lyapunov theory.

Theorem 3. *Consider tracking control issue (1) of the TMR. Suppose that an FCZNN model is polluted by the additive bounded error* $w_i(t)$ *with* $w_i(t) \le w$ *(constant or time-varying disturbance), where* w *is positive constant, starting from the arbitrary initial position* $q(0)$, *the steady-state tracking error of the FCZNN model (9) yields the following equality:*

$$\lim_{t\to+\infty}\left\|e(t)\right\|_2 < \sqrt{m}\left(\frac{w}{\gamma\kappa_3}\right)^{p_1/p}. \tag{29}$$

where all the parameters in the inequality have been defined before.

Proof: Provided that the additive disturbances exist in the FCZNN model, its ith dynamical subsystem corresponding to the error function in the FCZNN model is given by

$$\dot{e}_i(t) = -\gamma\phi(e_i(t)) + w_i \tag{30}$$

Similar to Theorem 1, a Lyapunov function is defined first to address the global convergence of the proposed FCZNN model.

$$L(t) = \frac{p}{p+p_1}e_i(t)^{\frac{p_1+p}{p}} \tag{31}$$

Obviously, $L(t)$ is an even function, $L(t) \ge 0$. Taking derivation for $L(t)$, we have

$$\begin{aligned} \dot{L}(t) &= e_i(t)^{p_1/p}\dot{e}_i(t) \\ &= \left[-\gamma\left(\kappa_1 e_i(t)^{p/p_1} + \kappa_2 e_i(t) + \kappa_3 e_i(t)^{p_1/p}\right) + w_i\right]e_i(t)^{p_1/p} \\ &= -\gamma\kappa_3\left(e_i(t)^{p_1/p} - \frac{w_i}{2\gamma\kappa_3}\right)^2 + \frac{{w_i}^2}{4\gamma\kappa_3} - \\ &\left(\gamma\kappa_1 e_i(t)^{(p^2+{p_1}^2)/pp_1} + \gamma\kappa_2 e_i(t)^{(p+p_1)/p}\right) \end{aligned} \tag{32}$$

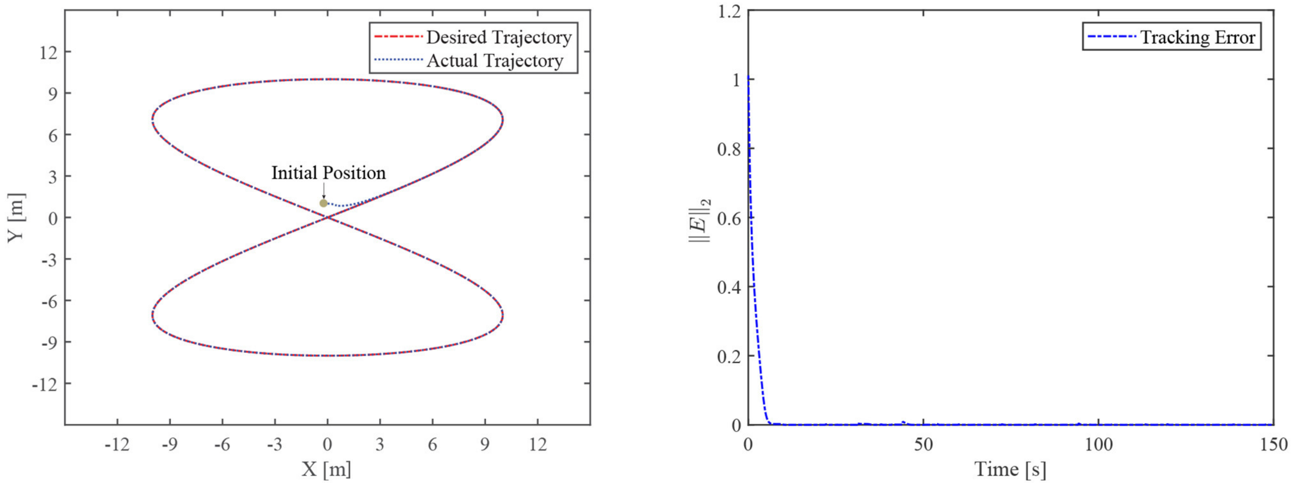


FIGURE 4
The tracking performance of the FCZNN without disturbance.

TABLE 2 The disturbances forms.

No.	Disturbance forms	Expression
1	Constant form	$w_i = 1$
2	Line form	$w_i = 0.01 * t$
3	Sine form	$w_i = sint$
4	Exponential decay form	$w_i = \exp(-t)$

Suppose that $e_i(t) \geq (w/\gamma\kappa_3)^{p/q}$, the first two terms hold $-\gamma\kappa_3\left(e_i(t)^{q/p} - \frac{w_i}{2\gamma\kappa_3}\right)^2 + \frac{w_i{}^2}{4\gamma\kappa_3} < 0$.

Based on this, we obtain the following analysis about Equation (32). There are two situations.

1) If solution error $e_i(t) \geq (w/\gamma\kappa_3)^{p/q}$ holds true, one can readily obtain that $\dot{L}(t) < 0$. In the sense of the Lyapunov theory, the system becomes stable gradually with time.

2) If solution error $e_i(t) < (w/\gamma\kappa_3)^{p/p_1}$ holds true, the sign of $\dot{L}(t)$ might be positive or negative. Even in the worst-case scenario, we consider $\dot{L}(t) > 0$, which indicates that $e_i(t)$ will increase; $(w_i/\gamma\kappa_3)^{p/p_1}$ does not exceed the upper bound $(w/\gamma\kappa_3)^{p/p_1}$ for $\dot{L}(t) < 0$ when $e_i(t) \geq (w_i/\gamma\kappa_3)^{p/p_1}$.

Recalling that $\|e(t)\|_2 = \sqrt{\sum_{i=1}^{m} e_i^2(t)}$, one can readily draw the conclusion that $\lim\limits_{t\to+\infty} \|e(t)\|_2 < \sqrt{m}\left(\frac{w}{\gamma\kappa_3}\right)^{p_1/p}$. The proof is thus completed. □

It is worth pointing out that Theorem 2 presents that the steady-state solution error can be arbitrarily small by increasing or reducing the fractional value.

Theorem 4. *In the case of $e_i(t) \geq (w_i/\gamma\kappa_3)^{p/p_1}$, starting from any initial value $q(0)$, the actual trajectory $q(t)$ tracks the desired position $q_d(t)$ in finite time T_f for the FCZNN model (9) with constant noises. T_f satisfies the following equality:*

$$T_f \leq \frac{1}{(a-1)\beta_2} \ln \frac{\beta_1 + \beta_2}{\beta_2 L^{1-a}(0) + \beta_1} \tag{33}$$

where the parameter in (33) is predefined in Theorem 2.

Proof: A Lyapunov function $L(t) = \left(e^+(t)\right)^2$ is defined; the derivate of $L(t)$ is demonstrated

$$\begin{aligned}\dot{L}(t) &= 2\dot{e}^+(t)e^+(t)\\ &= 2\left(-\gamma\Phi\left(e^+(t)\right) + w_i\right)e^+(t)\\ &= \Big(-2\gamma\kappa_1 L(t)^{(p+p_1)/2p_1} - 2\gamma\kappa_2 L(t) - 2\gamma\kappa_3 L(t)^{(p_1+p)/2p} +\\ &\quad 2w_i e^+(t)\Big)\end{aligned} \tag{34}$$

Then, $\dot{L}(t)$ is rewritten as $\dot{L}(t) = -(\beta_1 L^a(t) + \beta_2 L(t) + \beta_3 L^b(t)) + 2w_i e^+(t)$ Considering Theorem 3, if $e_i(t) \geq (w/\gamma\kappa_3)^{p/p_1}$ holds true (i.e., $w_i \leq \beta_3 e^+(t)^{p_1/p} \big/ 2$), one can have

$$\begin{aligned}2w_i e^+(t) &\leq \beta_3 e^+(t)^{(p+p_1)/p}\\ &= \beta_3 L(t)^{(p+p_1)/2p}\end{aligned} \tag{35}$$

Then, (34) is reformulated as

$$\begin{aligned}\dot{L}(t) &= -(\beta_1 L^a(t) + \beta_2 L(t) + \beta_3 L^b(t) + we^+(t))\\ &\leq (-\beta_1 L^a(t) - \beta_2 L(t) - \beta_3 L^b(t) + \beta_3 L^b(t))\\ &= -\beta_1 L^a(t) - \beta_2 L(t)\end{aligned} \tag{36}$$

□

Based on the discussion in Theorem 2, T_f satisfies (33). Then, the proof is completed.

4. Numerical experiments

The numerical experiments are conducted in this section to demonstrate the finite-time convergence and robustness of the

FCZNN model with disturbance considered. The CZNN is adopted for comparison.

During the initialization of the algorithm, the initial position vector is set to be $q(0) = q_d(0) + \Delta o$. The vector Δo is the off-set between an actual position and the desired position in the Cartesian space. $\Delta o = (0, 1, 0)$ is set in the simulation. The predefined design parameter is set to be $\gamma = 10$, and we keep $\kappa_i = 10$ for $i = 1, 2, 3$. Moreover, p and p_1 are set to be 9 and 3 separately. In the application, the TMR is applied to track an eight-shaped path. The reference trajectory for TMR is given by

$$\begin{cases} x_d = h_1 \sin(h_2 t), \\ y_d = h_1 \sin(h_3 t), \end{cases} t \in [0, T] \tag{37}$$

where $[h_1, h_2, h_3] = [10, 0.01, 0.05]$. Then, we have

Remark 1. *The scope of eight-shaped reference trajectory can be adjusted by changing the value of* h_1 *, that is,* $(x_d, y_d) \subset \{(x_d, y_d) \,|{-h_1} \le x_d \le h_1, -h_1 \le y_d \le h_1\}$.

$$\begin{cases} \dot{x}_d = h_1 h_2 \cos(h_2 t), \;\; \dot{y}_d = h_1 h_3 \cos(h_3 t), \\ \ddot{x}_d = -h_1 h_2^2 \sin(h_3 t), \;\; \ddot{y}_d = -h_1 h_3^2 \sin(h_3 t), \\ v_d = \sqrt{\dot{x}_d^2 + \dot{y}_d^2}, \\ \theta_d = \arctan 2(\dot{x}_d, \dot{y}_d). \end{cases} \tag{38}$$

The path-tracking task duration is set to be 150s as the initialization. Meanwhile, the general tracking error is expressed as the two norm of the error vector.

$$\|E\|_2 = \sqrt{(x - x_d)^2 + (y - y_d)^2 + (\theta - \theta_d)^2}. \tag{39}$$

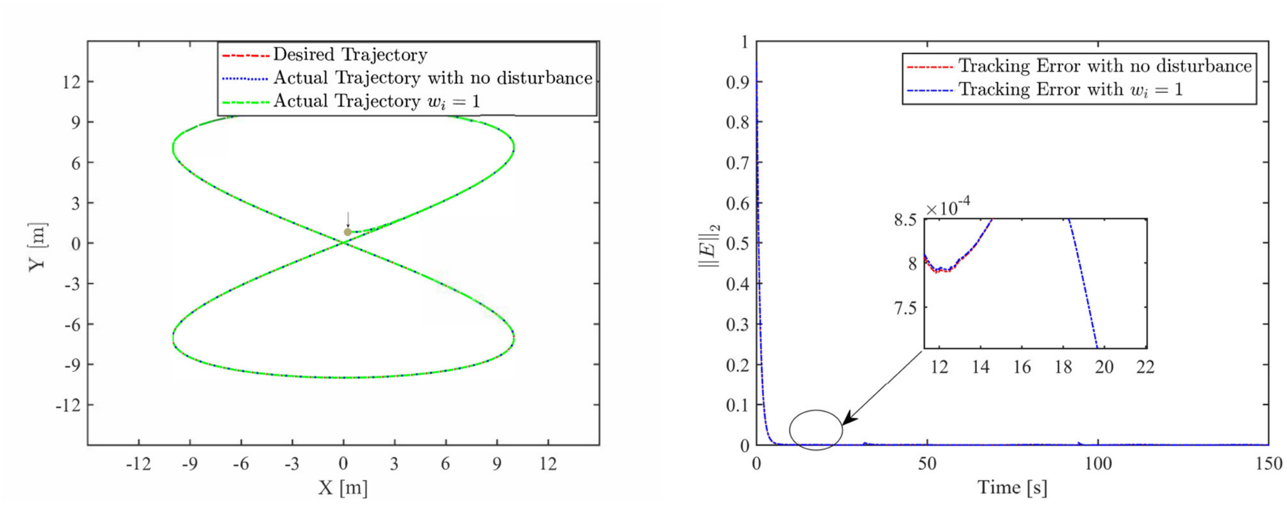


FIGURE 5
The tracking performance of the FCZNN model with disturbance $w_i = 1$.

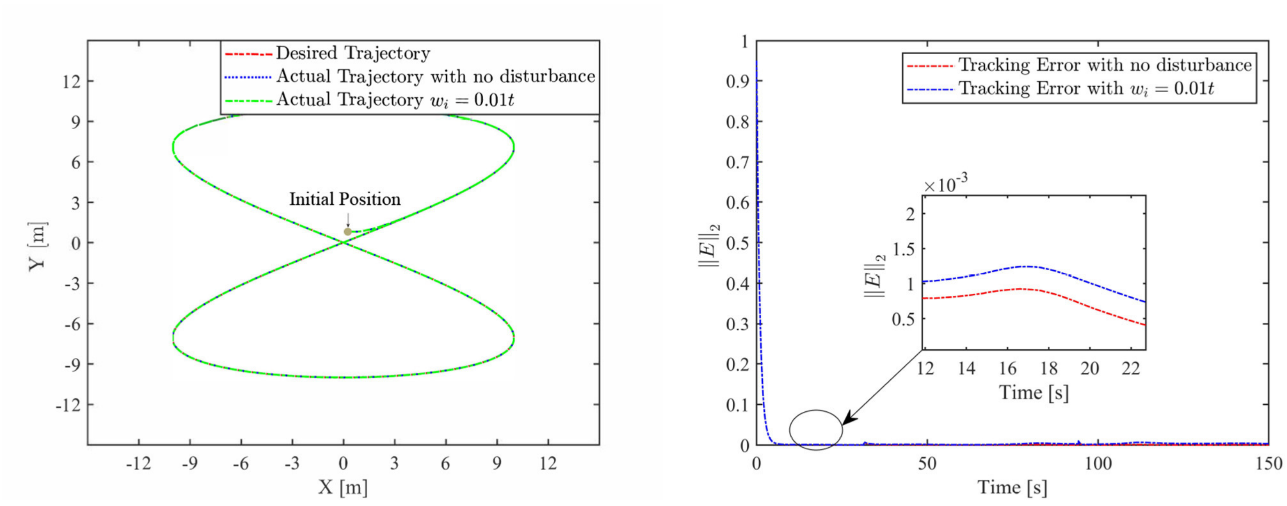


FIGURE 6
The tracking performance of FCZNN with disturbance $w_i = 0.01 * t$.

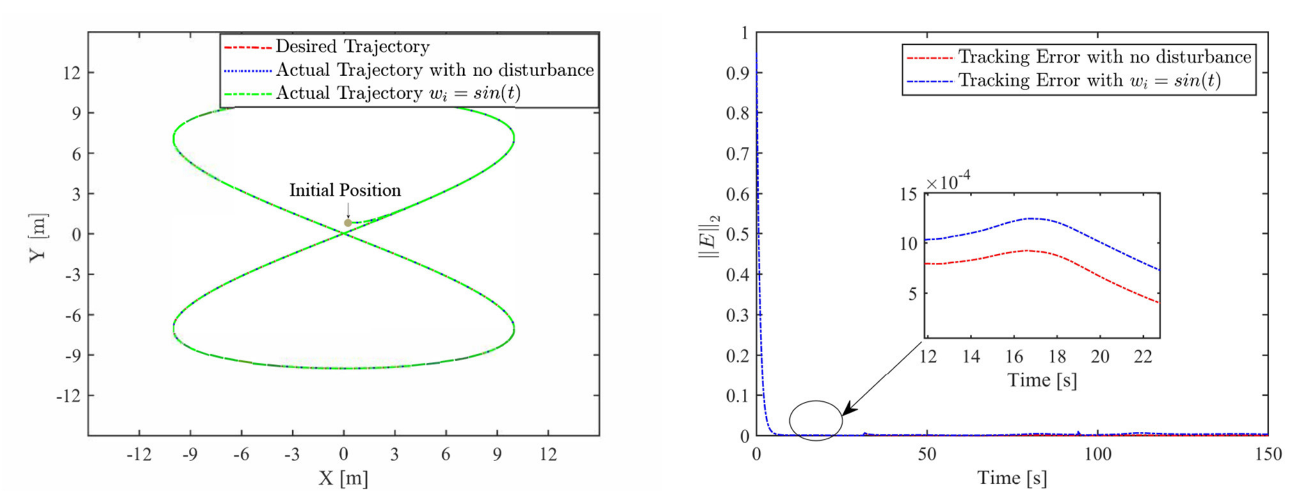


FIGURE 7
The tracking performance of the FCZNN model with disturbance $w_i = sin(t)$.

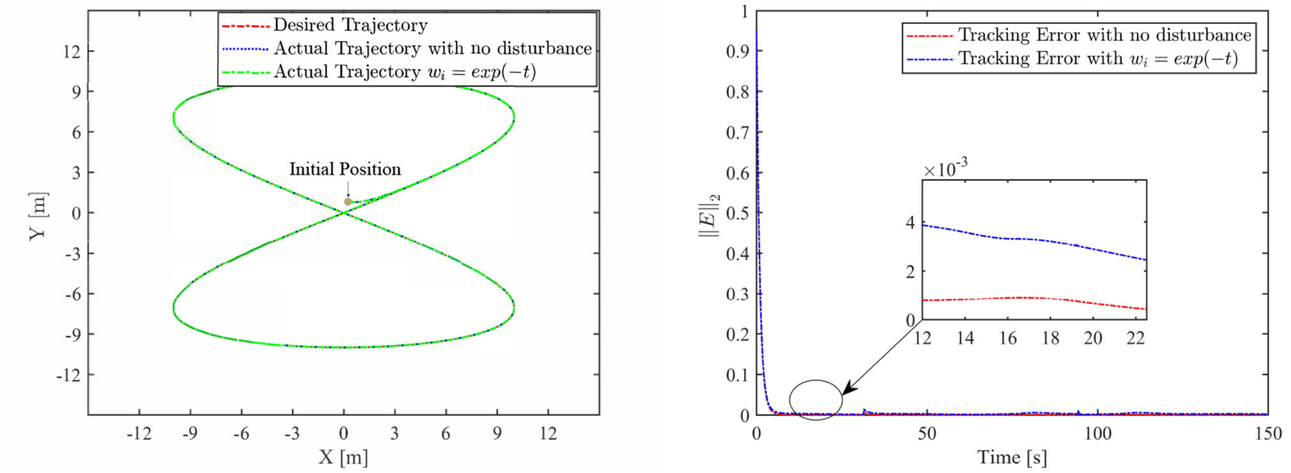


FIGURE 8
The tracking performance of the FCZNN model with disturbance $w_i = \exp(-t)$.

4.1. Finite-time convergence validation without disturbance

The simulative results of the FCZNN without disturbance are shown in Figure 4. Figure 4 presents the tracking performance for the TMR to track the eight-shaped path, which shows that the actual trajectory moves toward the desired trajectory and demonstrates the tracking error during the tracking task, validating the finite-time convergence with global stability. The tracking error decreases directly from the maximum value, which indicates that the error is related to the setting of the initial position because the error of the robot in the initial position is the maximum, consistent with the theoretical analysis.

4.2. Robustness verification

In general, disturbances are unavoidable for any electronic system and neural dynamics, mainly including internal and external disturbances. Internal disturbances are caused by hardware implementation off-set errors, which can be viewed as dynamic disturbances in linear or sinusoidal form. External disturbances are caused by instantaneous changes in power or external shock among other reasons, which can be regarded as the disturbance that disappears exponentially.

The disturbances considered in this study are shown in Table 2, including four different common disturbances.

Th motion results of the TMR tracking an eight-shaped path synthesized by the FCZNN model are shown in Figures 5–8.

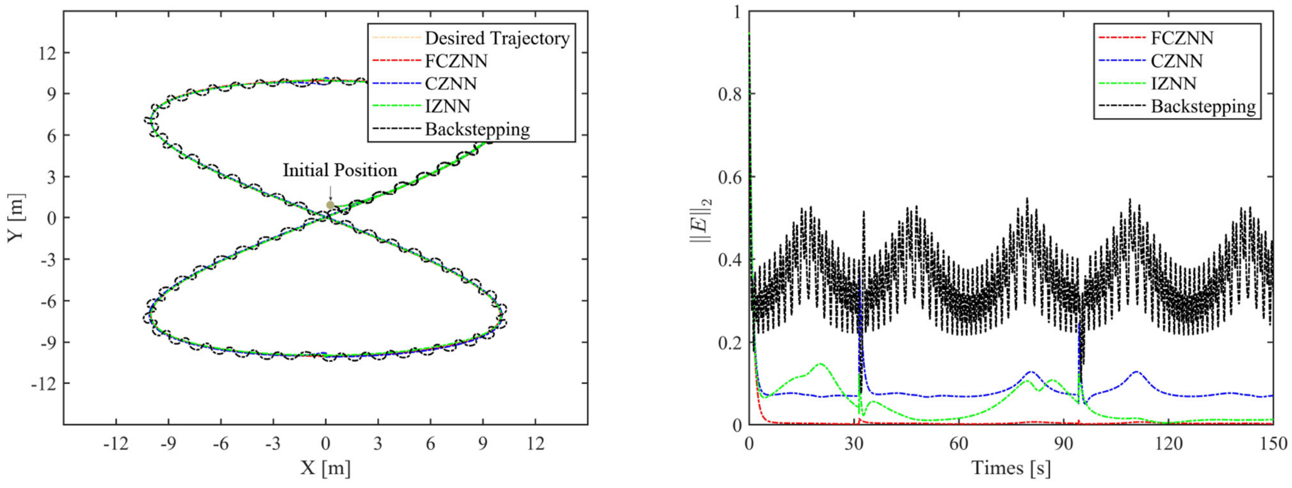


FIGURE 9
Simulated motion results with $w_i = 1$.

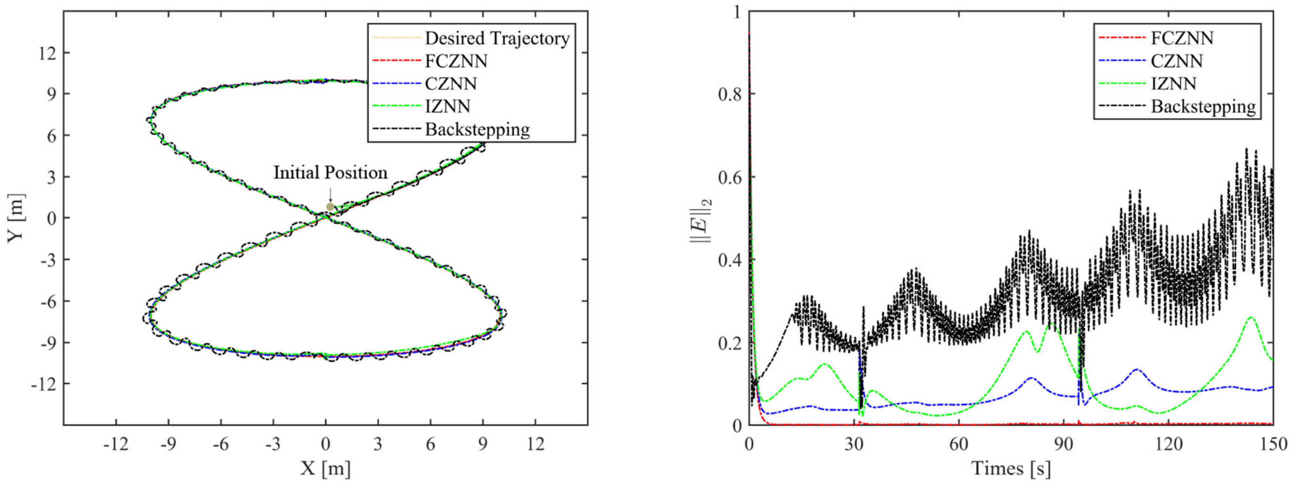


FIGURE 10
Simulated motion results with $w_i = 0.01t$.

Figure 5 shows that under constant value perturbations, the FCZNN model still has an excellent effect on the tracking error of the trajectory, indicating that it has a better suppression effect on constant value perturbations. Figures 6, 7 present that under linear or sine-form perturbations, there is still room for improvement in the suppression of the FCZNN model. Figure 7 illustrates that perturbations in the exponential decay form have a larger impact on the system at the moment they occur, unlike linear and sinusoidal perturbations.

Combing in the above figures, in the disturbance environment, the FCZNN model can still guarantee finite-time convergence. That is, in a disturbance environment, the TMR can still track the desired trajectory. Certainly, the convergence time is longer than that in Figure 4. The previous analysis illustrates that the tracking effect can be further enhanced by changing the parameters. In addition, we notice that the convergence time in the case of constant interference is longer than that in the case of time-varying disturbance. The upper limit of the time-varying disturbance is 1, and the time-varying disturbance is 0 at the beginning of the numerical experiment. Hence, the FCZNN model can track the desired trajectory faster.

4.3. Comparison with existing models

To verify the efficacy and superiority of the FCZNN model, comprehensive comparisons with existing neural network models are presented in this section, including the CZNN (Miao et al., 2015; Xiao et al., 2017) and integration-enhanced ZNN (IZNN) (Chen and Zhang, 2018; Xiao et al., 2019). Moreover, the classical backstepping control Hao et al. (2017) is introduced for comparison

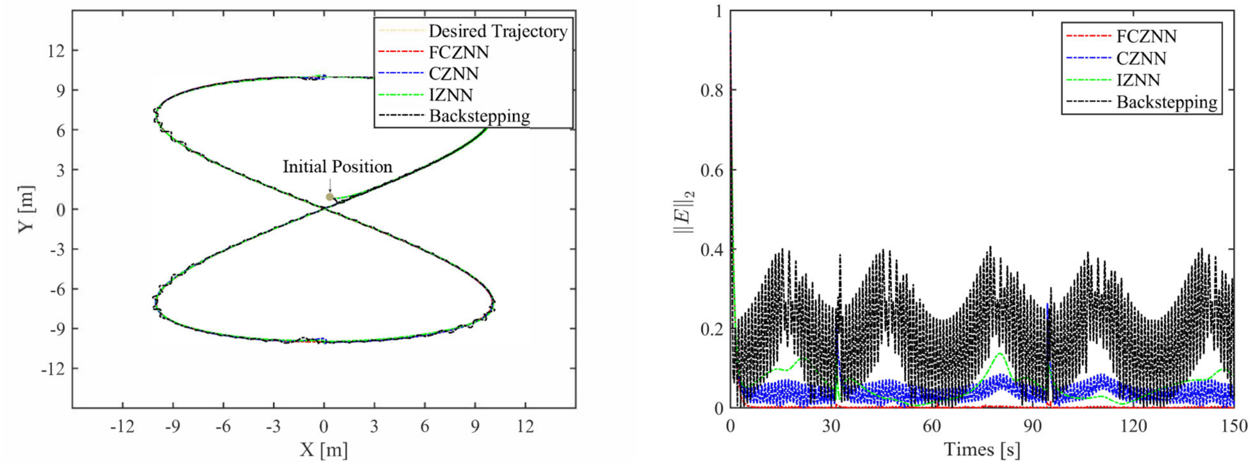


FIGURE 11
Simulated motion results with $w_i = sin(t)$.

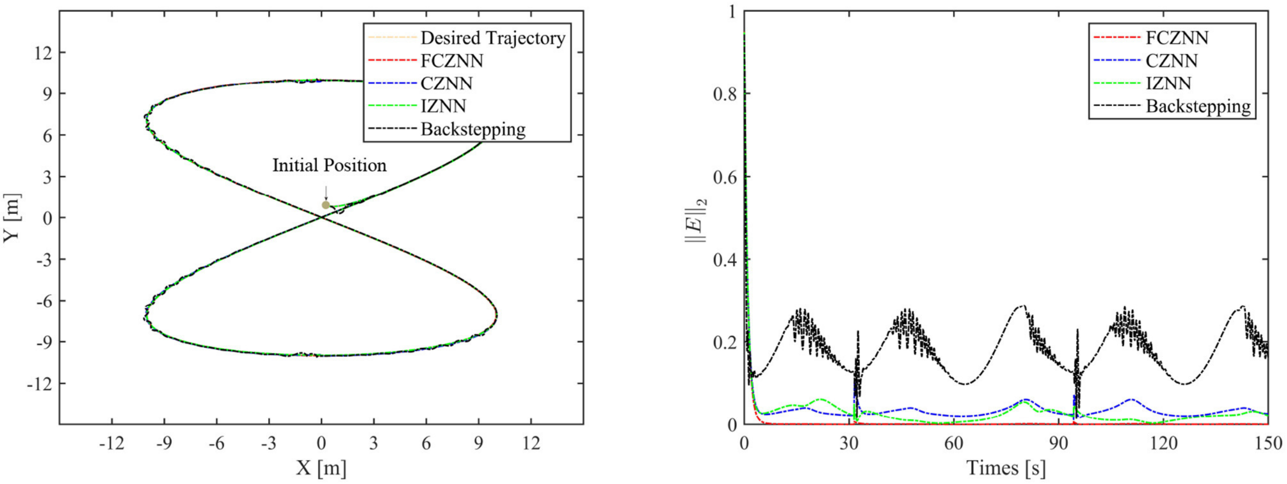


FIGURE 12
Simulated motion results with $w_i = \exp(-t)$.

as well. Figures 9–12 show the comparison results of various models with different disturbances. Clearly, all four methods are able to complete the task of trajectory tracking, but the quality differs considerably.

For solving the inverse kinematics problem of the mobile robot, the CZNN model with the disturbances can be depicted as the following dynamic equation:

$$A(t)u(t) - \dot{q}_d(t) = -\gamma e(t) + W(t) \tag{40}$$

The convergence feature of the CZNN model without disturbance has been investigated broadly and is neglected in this study. Without loss of generality, parameters γ and κ_i for $i = 1, 2, 3$ are kept the same.

The blue line in Figures 9–12 demonstrates the tracking performance of the CZNN model and its tracking error, showing that this model is sensitive to disturbances, especially the three time-varying disturbances. Figure 9 shows that the maximum tracking error of this model is much higher than that of the FCZNN and IZNN models. Generally, the tracking error of the CZNN model does not converge to be 0 during the entire tracking duration. Therefore, the CZNN model is not suitable for application in the disturbance environment.

The IZNN model has been presented and investigated as an alternative for solving the inverse kinematics problem of mobile robot manipulators; this model with disturbances can be depicted as the following dynamic equation:

$$A(t)u(t) - \dot{q}_d(t) = -\gamma e(t) - \lambda \int \Phi e(t)dt + W(t) \tag{41}$$

Readily, the simulation results present that the performance of the IZNN model is enhanced compared to that of the CZNN model. Figures 9–11 present that the IZNN model is nonsensitive to constant and exponential decay disturbances, but it cannot deal with sine or linear disturbances effectively. It does not meet our requirements.

Backstepping control is the classical method for solving the inverse kinematics problem of the mobile robot. However, the simulation results present its failure in achieving satisfactory results in an interference environment. Specifically, its tracking trajectory is not smooth, not to mention its tracking error. Details about backstepping control will, therefore, not be discussed in the paper.

Figures 9–12 illustrate that the proposed FCZNN model exhibits anti-disturbance performance with four common forms of disturbances suppressed for solving the inverse kinematics problem of the TMR compared with the existing two models and backstepping control. In addition, comparisons with other models or methods with the corresponding results shown in Figure 9 substantiate the robust property and finite convergence of the proposed FCZNN model, which are absent in both the CZNN and IZNN models.

Based on the above simulation results and analysis, we can draw the conclusion that the proposed FCZNN model has excellent and inherent noise and disturbance canceling ability accompanied by finite-time convergence, which enables it to be more suitable for practical applications of the TMR with noises and disturbances.

5. Conclusion

An FCZNN model was proposed in this study as a solution to the TMR tracking control . Different from the CZNN model, a new activation function was incorporated with the FCZNN model. Some theorems of finite-time convergence and strong robustness were mathematically validated. Simulation experiments were conducted to verify the superiority and effectiveness of the proposed FCZNN model in comparison with the CZNN, IZNN, and backstepping control. Furthermore, the application to TMR kinetic control presented its practical significance.

Future work lies in extending the kinematic analysis by considering multiple physical constraints and developing a complete experimental environment equipped with the real TMR for practical application of the FCZNN model. The extension of the FCZNN model to other similar mechanisms is an interesting, open, and challenging future direction for this research. Moreover, developing ZNN models that consider obstacle avoidance and saturation constraints to enable the TMR with active obstacle avoidance or developing novel saturation-allowed activation functions to adapt to practical requirements is an interesting research direction.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

YC and BL: data curation and validation. YC, BL, and JP: conceptualization and formal analysis. YC: methodology, software, and writing—original draft. JP: supervision. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ahmed, S., Wang, H., and Tian, Y. (2021). Adaptive fractional high-order terminal sliding mode control for nonlinear robotic manipulator under alternating loads. *Asian J. Control* 23, 1900–1910. doi: 10.1002/asjc.2354

Bistron, M., and Piotrowski, Z. (2021). Artificial intelligence applications in military systems and their influence on sense of security of citizens. *Electronics* 10, 871. doi: 10.3390/electronics10070871

Chen, D., Cao, X., and Li, S. (2021). A multi-constrained zeroing neural network for time-dependent nonlinear optimization with application to mobile robot tracking control. *Neurocomputing* 460, 331–344. doi: 10.1016/j.neucom.2021.06.089

Chen, D., Li, S., and Wu, Q. (2020). A novel supertwisting zeroing neural network with application to mobile robot manipulators. *IEEE Trans. Neural Netw. Learn. Syst.* 32, 1776–1787. doi: 10.1109/TNNLS.2020.2991088

Chen, D., and Zhang, Y. (2018). Robust zeroing neural-dynamics and its time-varying disturbances suppression model applied to mobile robot manipulators. *IEEE Trans. Neural Netw. Learn. Syst.* 29, 4385–4397. doi: 10.1109/TNNLS.2017.2764529

Ding, L., Li, S., Gao, H., Chen, C., and Deng, Z. (2018). Adaptive partial reinforcement learning neural network-based tracking control for wheeled mobile robotic systems. *IEEE Trans. Syst. Man Cybern. Sys.* 50, 2512–2523. doi: 10.1109/TSMC.2018.2819191

Fang, Y., Huang, Z., Pu, J., and Zhang, J. (2022). Auv position tracking and trajectory control based on fast-deployed deep reinforcement learning method. *Ocean Eng.* 245, 110452. doi: 10.1016/j.oceaneng.2021.110452

Gao, M.-M., Jin, X.-Z., and Ding, L.-J. (2022). Robust adaptive backstepping intsm control for robotic manipulators based on elm. *Neural Comput. Appl.* 34, 5029–5039. doi: 10.1007/s00521-021-05824-y

Gerontitis, D., Behera, R., Tzekis, P., and Stanimirović, P. (2022). A family of varying-parameter finite-time zeroing neural networks for solving time-varying sylvester equation and its application. *J. Comput. Appl. Math.* 403, 113826. doi: 10.1016/j.cam.2021.113826

Gu, Q., Bai, G., Meng, Y., Wang, G., Zhang, J., Zhou, L., et al. (2021). Efficient path tracking control for autonomous driving of tracked emergency rescue robot under 6G network. *Wirel. Commun. Mob. Comput*. 2021, 1–9. doi: 10.1155/2021/5593033

Hao, Y., Wang, J., Chepinskiy, S. A., Krasnov, A. J., and Liu, S. (2017). "Backstepping based trajectory tracking control for a four-wheel mobile robot with differential-drive steering," in *2017 36th Chinese Control Conference (CCC)* (Dalian: IEEE), 918–4923. doi: 10.23919/ChiCC.2017.8028131

Hu, Z., Li, K., Li, K., Li, J., and Xiao, L. (2020). Zeroing neural network with comprehensive performance and its applications to time-varying lyapunov equation and perturbed robotic tracking. *Neurocomputing* 418, 79–90. doi: 10.1016/j.neucom.2020.08.037

Ji, M., Sun, Z., Wang, J., and Chen, Q. (2002). "Robust backstepping control of tracked mobile robot," in *Mobile Robots XVI*, volume 4573 (Bellingham, WA: SPIE), 235–243. doi: 10.1117/12.457448

Jin, J., and Qiu, L. (2022). A robust fast convergence zeroing neural network and its applications to dynamic sylvester equation solving and robot trajectory tracking. *J. Franklin Inst*. 359, 3183–3209. doi: 10.1016/j.jfranklin.2022.02.022

Lara-Molina, F. A., and Dumur, D. (2021). A fuzzy approach for the kinematic reliability assessment of robotic manipulators. *Robotica* 39, 2095–2109. doi: 10.1017/S0263574721000187

Li, J., Wang, J., Peng, H., Hu, Y., and Su, H. (2022). Fuzzy-torque approximation-enhanced sliding mode control for lateral stability of mobile robot. *IEEE Trans. Syst. Man Cybern. Syst*. 52, 2491–2500. doi: 10.1109/TSMC.2021.3050616

Luo, J., Yang, H., Yuan, L., Chen, H., and Wang, X. (2022). Hyperbolic tangent variant-parameter robust znn schemes for solving time-varying control equations and tracking of mobile robot. *Neurocomputing* 510, 218–232. doi: 10.1016/j.neucom.2022.08.066

Ma, Z., Yu, S., Han, Y., and Guo, D. (2021). Zeroing neural network for bound-constrained time-varying nonlinear equation solving and its application to mobile robot manipulators. *Neural Comput. Appl*. 33, 14231–14245. doi: 10.1007/s00521-021-06068-6

Miao, P., Shen, Y., Huang, Y., and Wang, Y.-W. (2015). Solving time-varying quadratic programs based on finite-time zhang neural networks and their application to robot tracking. *Neural Comput. Appl*. 26, 693–703. doi: 10.1007/s00521-014-1744-4

Rawat, R., Rajawat, A. S., Mahor, V., Shaw, R. N., and Ghosh, A. (2021). "Surveillance robot in cyber intelligence for vulnerability detection," in *Machine Learning for Robotics Applications*, eds M. Bianchini, M. Simic, A. Ghosh, and R. N. Shaw (Singapore: Springer), 107–123. doi: 10.1007/978-981-16-0598-7_9

Sabiha, A. D., Kamel, M. A., Said, E., and Hussein, W. M. (2022). Ros-based trajectory tracking control for autonomous tracked vehicle using optimized backstepping and sliding mode control. *Rob. Auton. Syst*. 152, 104058. doi: 10.1016/j.robot.2022.104058

Šegota, S. B., Anđelić, N., Mrzljak, V., Lorencin, I., Kuric, I., Car, Z., et al. (2021). Utilization of multilayer perceptron for determining the inverse kinematics of an industrial robotic manipulator. *Int. J. Adv. Robot. Syst*. 18, 1729881420925283. doi: 10.1177/1729881420925283

Sun, Z., Wang, G., Jin, L., Cheng, C., Zhang, B., Yu, J., et al. (2022). Noise-suppressing zeroing neural network for online solving time-varying matrix square roots problems: a control-theoretic approach. *Expert Syst. Appl*. 192, 116272. doi: 10.1016/j.eswa.2021.116272

Truong, T. N., Vo, A. T., and Kang, H.-J. (2021). A backstepping global fast terminal sliding mode control for trajectory tracking control of industrial robotic manipulators. *IEEE Access* 9, 31921–31931. doi: 10.1109/ACCESS.2021.3060115

Wang, Y., Huang, S., Wang, Z., Hu, R., Feng, M., Du, P., et al. (2022). Design and experimental results of passive iusbl for small auv navigation. *Ocean Eng*. 248, 110812. doi: 10.1016/j.oceaneng.2022.110812

Xiao, L., Dai, J., Jin, L., Li, W., Li, S., Hou, J., et al. (2019). A noise-enduring and finite-time zeroing neural network for equality-constrained time-varying nonlinear optimization. *IEEE Trans. Syst. Man Cybern. Syst*. 51, 4729–4740. doi: 10.1109/TSMC.2019.2944152

Xiao, L., Liao, B., Li, S., Zhang, Z., Ding, L., Jin, L., et al. (2017). Design and analysis of ftznn applied to the real-time solution of a nonstationary lyapunov equation and tracking control of a wheeled mobile manipulator. *IEEE Trans. Ind. Inform*. 14, 98–105. doi: 10.1109/TII.2017.2717020

Yan, X., Liu, M., Jin, L., Li, S., Hu, B., Zhang, X., et al. (2019). New zeroing neural network models for solving nonstationary sylvester equation with verifications on mobile manipulators. *IEEE Trans. Ind. Inform*. 15, 5011–5022. doi: 10.1109/TII.2019.2899428

Yin, X., Pan, L., and Cai, S. (2021). Robust adaptive fuzzy sliding mode trajectory tracking control for serial robotic manipulators. *Robot. Comput. Integr. Manuf*. 72, 101884. doi: 10.1016/j.rcim.2019.101884

Zhang, M., and Zheng, B. (2022). Accelerating noise-tolerant zeroing neural network with fixed-time convergence to solve the time-varying sylvester equation. *Automatica* 135, 109998. doi: 10.1016/j.automatica.2021.109998

Zhang, Y., Jiang, D., and Wang, J. (2002). A recurrent neural network for solving sylvester equation with time-varying coefficients. *IEEE Trans. Neural Netw*. 13, 1053–1063. doi: 10.1109/TNN.2002.1031938

frontiers | Frontiers in Neurorobotics

TYPE Original Research
PUBLISHED 25 September 2023
DOI 10.3389/fnbot.2023.1278986

OPEN ACCESS

EDITED BY
Long Jin,
Lanzhou University, China

REVIEWED BY
Miaomiao Ma,
North China Electric Power University, China
Tao Hou,
Jilin University, China
Jing Wang,
Henan Polytechnic University, China
Jia Yang,
Jilin Jianzhu University, China

*CORRESPONDENCE
Yunqing Liu
✉ mzliuyunqing@163.com

RECEIVED 17 August 2023
ACCEPTED 07 September 2023
PUBLISHED 25 September 2023

CITATION
Yan F, Liu L, Ding X, Zhang Q and Liu Y (2023) Monocular catadioptric panoramic depth estimation via improved end-to-end neural network model.
Front. Neurorobot. 17:1278986.
doi: 10.3389/fnbot.2023.1278986

Monocular catadioptric panoramic depth estimation via improved end-to-end neural network model

Fei Yan[1,2], Lan Liu[1], Xupeng Ding[1], Qiong Zhang[1,2] and Yunqing Liu[1,2]*

[1]School of Electronic Information Engineering, Changchun University of Science and Technology, Changchun, China, [2]New Technology Research Department, Jilin Provincial Science and Technology Innovation Center of Intelligent Perception and Information Processing, Changchun, China

In this paper, we propose a monocular catadioptric panoramic depth estimation algorithm based on an improved end-to-end neural network model. First, we use an enhanced concentric circle approximation unfolding algorithm to unfold the panoramic images captured by the catadioptric panoramic camera and then extract the effective regions. In addition, the integration of the Non-local attention mechanism is exploited to improve image understanding. Finally, a depth smoothness loss strategy is implemented to further enhance the reliability and precision of the estimated depths. Experimental results confirm that this refined algorithm is capable of providing highly accurate estimates of object depth.

KEYWORDS

catadioptric panoramic camera, panoramic image, depth estimation, attention model, depth smoothness loss

1. Introduction

Traditional camera systems are often limited by their narrow field of view, a problem that is currently being alleviated by the introduction of panoramic cameras (Svoboda et al., 1998). There are four main types of panoramic vision cameras: pan-tilt rotating, fisheye lens, multi-camera stitching, and catadioptric. In particular, a catadioptric panoramic camera uses a special type of mirror, called a catadioptric mirror, to direct light from different angles onto a single image sensor, thus capturing panoramic images (Jaramillo et al., 2016). Consisting mainly of a convex reflecting mirror, an imaging lens, and a photosensitive component (Baker and Nayar, 1998, 1999), the catadioptric panoramic camera avoids the complicated designs associated with optical lens structures and solves the problem of image distortion (Liu et al., 2016). Additionally, it eliminates the call for image stitching, thus affirming the real-time capture of a 360° panoramic view.

The rapid growth of visual systems research has increasingly made panoramic vision systems a critical point of interest for researchers in related fields. This technology finds its extensive applications in areas such as robotic navigation, Internet of Things (IoT), and autonomous driving (Yamazawa et al., 1995; Liu and Liang, 2013; Khurana and Armenakis, 2018). Panoramic vision systems are designed to capture a 360° view of the environment (Nichols et al., 2010). In the field of depth estimation in panoramic vision, a depth value is computed for each pixel in an image to facilitate the approximation of distances between objects in the scene and the camera itself. Two main approaches have dominated the research field of image depth estimation: supervised and unsupervised learning.

Supervised learning is performed on datasets that are comprehensively labeled with critical depth information, providing an effective method for monocular depth estimation (Eigen et al., 2014). A unique image reconstruction loss function is incorporated to assess the disparity between the generated depth map and the input image (Li et al., 2017), thereby supporting the network's learning of image depth information. In addition, data augmentation techniques are used to amplify and transform the training data, thus diversifying the network training samples and effectively increasing the network's generalization capacity (Eldesokey et al., 2020; Kusupati et al., 2020). Despite their proven ability to deliver high-quality depth estimation results, these methods are highly dependent on the considerable time and skill of the personnel responsible for the annotation process, making the potential occurrence of annotation errors or inconsistencies virtually unmanageable.

With the advancement of deep learning techniques, unsupervised end-to-end depth estimation methods have become one of the research hotspots. End-to-end neural network models can complete the entire process from input to output without the need for human intervention at intermediate steps. These models fall into two categories: the first assimilates learning through stereo matching techniques; the second exploits the displacement between successive frames to infer the depth data associated with objects in the scene (Garg et al., 2016). The use of unlabeled monocular video sequences as network inputs to train convolutional neural networks (CNNs) in an unsupervised approach has enabled depth estimation models to be independent of labeled depth information datasets (Zhou et al., 2017). This method has expanded the potential application scenarios of depth estimation models. However, a limitation of this method is the relatively lower precision of depth estimation. Consequently, various methodologies have been adopted to enhance the performance and robustness of depth estimation. These include the employment of a reconstruction image loss function to improve the consistency between left and right disparity maps (Godard et al., 2017), and the integration of three-dimensional geometric constraints to constrain unsupervised learning of depth (Mahjourian et al., 2018). By using binary depth classification during the training process, it is possible to quickly predict nearby objects (Badki et al., 2020). In addition, even with relatively coarse quantization of depth estimation, a high level of accuracy can be maintained. To tackle the prevalent issue of unsupervised scale, joint training of monocular depth estimation and stereo visual odometry is executed through the utilization of depth information derived from stereo images relative to the motion between them (Zhan et al., 2018). Unsupervised learning methods can automatically discover the depth structure within images without the need for any manual intervention. However, the complex phenomena in real outdoor scenes, such as lighting variations, occlusions, etc., pose potential challenges to image depth estimation. During image depth estimation, these factors can potentially lead to issues such as the loss of fine details in the predicted depth map and lower accuracy in the depth map, thereby preventing the acquisition of accurate depth information.

This paper presents a novel approach to depth estimation in panoramic images using a catadioptric panoramic camera. The unique design of this camera facilitates real-time monitoring of the environment in a 360° fashion and mitigates the challenges of distortion and missing patches encountered by multiple camera systems, ultimately reducing costs. The unsupervised end-to-end depth estimation method proposed herein systematically addresses the challenge of insufficiently accurate fine detail prediction often seen in existing models. With this goal in mind, our approach incorporates a Non-local attention mechanism to capture intricate contextual dependencies within images. Additionally, we introduce a depth smoothing loss to increase the accuracy and efficiency of our depth estimation algorithm.

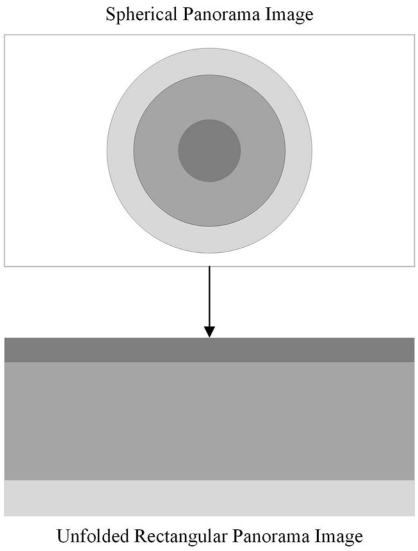


FIGURE 1
Panorama expansion.

2. Proposed method

2.1. Catadioptric panoramic camera image preprocessing

The imaging principles, manufacturing costs, and complexity of various curved reflecting mirrors are all factors to consider when selecting reflecting mirrors for a catadioptric panoramic imaging system. A hyperbolic mirror can capture images within a broader range and it offers the advantage of lower production costs. Therefore, in this paper, hyperbolic mirrors are selected as the reflecting elements for the catadioptric panoramic imaging system. Due to the special characteristics of the imaging principle of the catadioptric panoramic camera, the panoramic image captured by the catadioptric panoramic camera has a large distortion. To solve this problem, it is necessary to expand the panoramic image into a two-dimensional rectangular image, so that each pixel in the panoramic image corresponds to a position in the expanded image. This is called panorama expansion. As shown in Figure 1.

The traditional catadioptric panorama is usually expanded by the concentric circle approximate expansion algorithm, but

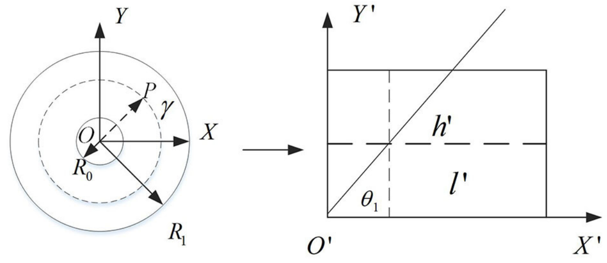


FIGURE 2
The principle of the improved concentric circle approximation unfolding algorithm.

the distortion of the expanded image is obvious, which will affect the subsequent processing of depth estimation. Aiming at this problem, this paper improved the concentric circle approximate expansion algorithm to reduce the distortion degree of the expanded image. Figure 2 shows the principle of the improved algorithm.

A rectangular coordinate system with the center of the catadioptric panorama as the origin O and the horizontal and vertical directions as the X-axis and Y-axis. The dashed line in the right plot of Figure 2 is shown. Let the ring represented by the dotted line be its panorama expansion. Thus, after the panorama expansion, the length and width of the 2D rectangle can be obtained. As shown in the following formula:

$$W = l' = 2\pi r, H = h' = r - R_0 \tag{1}$$

A ray passing through the center point O intersects a circular ring represented by a dashed line at a point $P(x_1, y_1)$. After unfolding, the angle between the ray and the X-axis is denoted by θ_1, so:

$$\theta_1 = \frac{1}{R_1} \tag{2}$$

Using ray OP as polar axis, rotate 360° around pole point O. By calculation, all the pixel values on the circumference of the circle can be obtained and arranged in a certain order. The calculation formula is:

$$\begin{cases} \rho = H + R_0 \\ x = \rho\cos(\theta_1) + u_0 \\ y = \rho\sin(\theta_1) + v_0 \end{cases} \tag{3}$$

As shown in Figure 3, the interpolation process of the panoramic image is depicted. In Figure 3A, the red dashed lines represent the pixel values after unfolding the catadioptric panoramic image. In Figure 3B, the black dots represent the inserted pixel values. In Figure 3A, the longest side of the trapezoid corresponds to the region farthest away from the center of the catadioptric panoramic image. Based on this longest side, construct a two-dimensional rectangle, ensuring that each row has the same number of interpolated pixels as the length of the longest side. If the longest side of the trapezoidal image has m pixels, and the shortest side has n pixels, then the shortest side needs to insert m-n pixels to ensure that each row in the two-dimensional rectangle has the same number of pixels as the longest side. When performing the interpolation process, the first pixel $x_i, i = 1, 2, \cdots m$ of the shortest side of the trapezoidal image should be placed at the first position of the corresponding side of the rectangle. Since we need to insert $m - n$ pixels on the shortest side, this means that between adjacent pixels, we will need to insert $(m - n)/n$ pixels to maintain the required consistency in the interpolation process. By using interpolation, we insert n interpolated pixel values between adjacent pixels of the catadioptric panoramic image's shortest side. This process is performed consistently for each row, resulting in the final rectangular unfolded image of the catadioptric panoramic view. Finally, by using interpolation, we insert $(m - n)/n$ pixel values between adjacent pixels of the longest side of the catadioptric panoramic image, resulting in the final rectangular unfolded image of the catadioptric panoramic view.

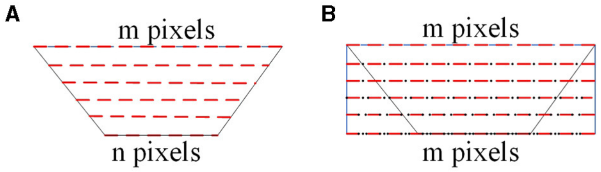


FIGURE 3
The principle of interpolation for catadioptric panoramic images. **(A)** is unfolding of the catadioptric panoramic image, and **(B)** is unfolding image after interpolation.

As shown in Figure 4, the simulation results from both methods indicate that the improved method exhibits significantly better real-time performance compared to the traditional method. Comparing the unfolded images in Figures 4B, C, the improved concentric circle approximation unfolding algorithm shows less distortion and more accurately reproduces the original scene.

The unfolded panoramic image contains complete 360° panoramic information of the scene. In practice, only the part of the image directly in front of the object is needed. Therefore, it is necessary to extract the relevant region from the unfolded panoramic image effectively. From the unfolded image in Figure 4C, it can be observed that the frontal view of the vehicle is located on the left side of the unfolded image, and the height of the vehicle's top part is approximately one-third of the entire image height. Therefore, the effective region of interest lies within the left-to-right half of the unfolded image and within the top-to-bottom two-thirds of the unfolded image height. As shown in Figure 5, this effective region is the crucial area for subsequent depth estimation. In fact, this can reduce the computational load, improve detection speed, and even eliminate some false positives.

2.2. Improved unsupervised monocular depth estimation model

To address the challenge of difficult annotation in supervised learning methods, this paper adopts an unsupervised learning

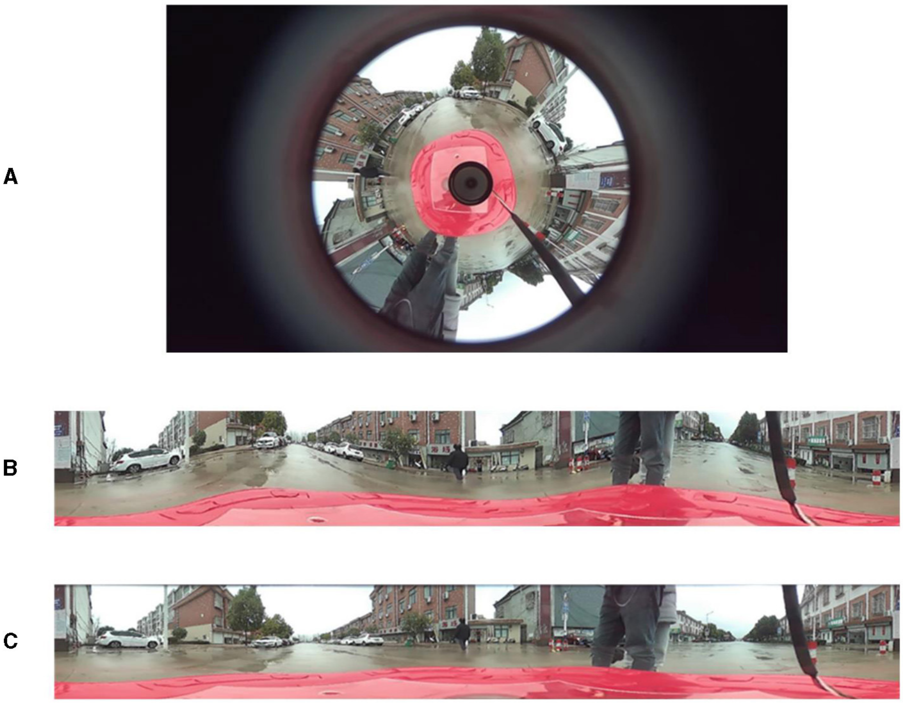


FIGURE 4
Unfolding of the catadioptric panoramic image. **(A)** is the panoramic image, **(B)** is the unfolded image using the traditional method, and **(C)** is the unfolded image using the improved method.

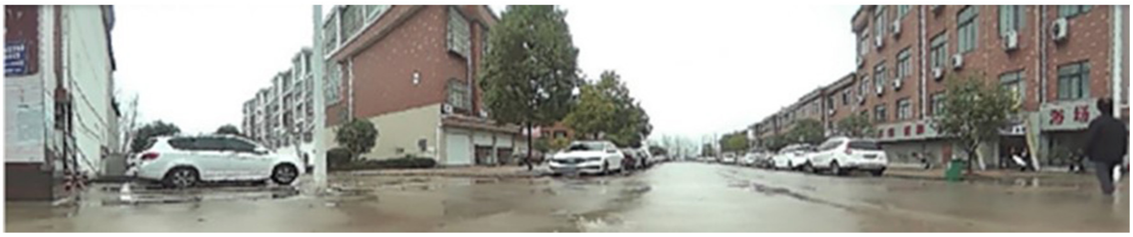


FIGURE 5
Effective region extraction.

approach for image depth estimation. A common problem with unsupervised end-to-end depth estimation methods is the lack of accuracy in predicting fine details. Existing unsupervised image depth estimation methods pay limited attention to the influence of the spatial context of the image on the depth information. Therefore, this paper proposes an improvement to a novel unsupervised learning algorithm framework by incorporating the Non-local attention mechanism module into the network structure of the encoder and decoder. This helps the network to perform adaptive contextual modeling for different regions in the image. This method enables the network to better comprehend various objects, backgrounds, and textures present in the image, thereby enhancing its understanding and representation capabilities of the image content.

As shown in Figure 6, this is the improved unsupervised learning depth estimation network model. The network is based on an end-to-end encoder-decoder framework, allowing it to perform depth estimation on images at multiple scales. In order to better capture contextual information in the image, a Non-local operation attention mechanism module is incorporated into the network framework. In each layer of the encoder, the Non-local operation attention mechanism module is used as the second operation and employs convolution with a stride of 2. The network architecture consists of three parts: an encoder, a decoder, and a Non-local operation module. The encoder is used for feature extraction, responsible for converting the input image into high-dimensional feature vectors. Convolutional neural networks (CNNs) are commonly employed to implement the encoder part. The decoder is used for depth estimation, and its main role is to decode the feature vectors extracted by the encoder into a depth map. The Non-local operation module is used to extract contextual information from the image and enables global interaction in the

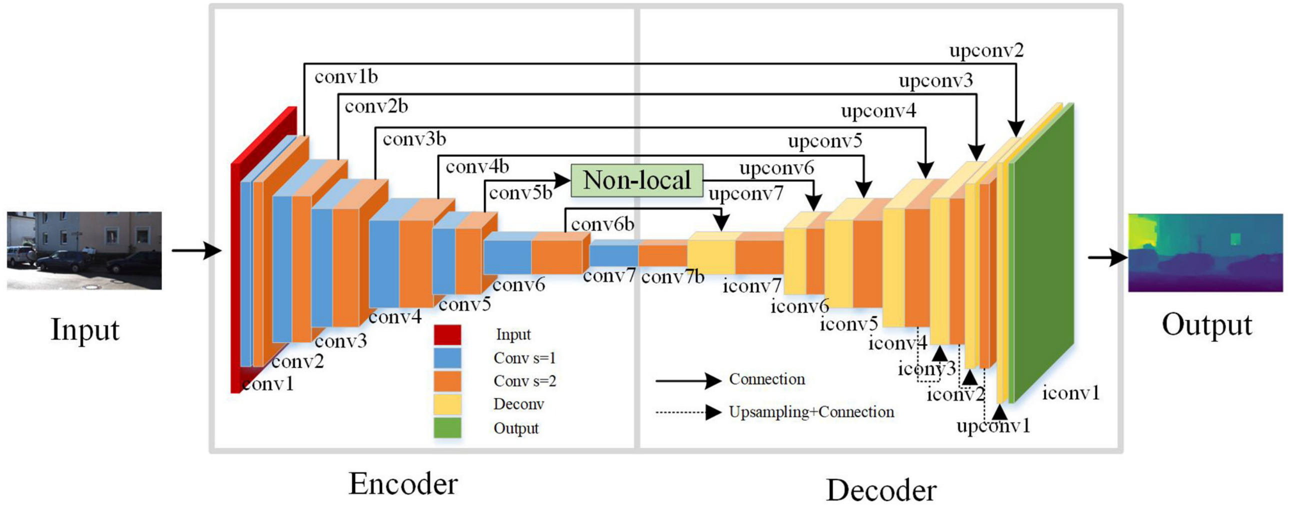


FIGURE 6
Unsupervised learning-based image depth estimation model.

spatial dimension of the input feature map. This allows for the fusion of global contextual information, helping the model to better understand the relationships between objects in the image, leading to improved performance.

In this paper, the Disp Net framework (Mayer et al., 2016) is used to design the structure of the encoder. And, by combining the long-range skip connections and the Non-local operation, the network's expressive power is enhanced to obtain more accurate depth maps. Convolutional layers are mainly used for feature extraction in neural networks. Activation layers introduce nonlinearity into the neural network, which is essential for the network to learn complex and nonlinear patterns in the data. Pooling layers play a crucial role in reducing the spatial dimensions of the feature maps, which can help in reducing the computational load and the number of parameters in the network. In this network architecture, except for the output layer, ReLU activation functions are used after all the convolutional layers. That's because this activation function has advantages such as fast computation, ease of optimization, and avoidance of the vanishing gradient problem. This design strategy helps to enhance the performance and stability of the network, making the depth estimation model more reliable and practical.

In the encoder, using convolutional operations with a stride of 2 is intended to extract features more efficiently. This convolutional operation helps to reduce the size of feature maps, increase the receptive field, and decrease the number of channels, thereby reducing computational complexity, lowering memory consumption, and improving the computational efficiency of the network. Increasing the receptive field helps the network to better understand the contextual information in the input data, thereby improving the prediction accuracy of the network. Reducing the number of channels in the feature maps helps to lower the dimensionality of the data, leading to reduced computational and storage costs. By combining these operations in the encoder, the performance and efficiency of the neural network can be effectively optimized. Finally, the predicted depth values are constrained using the function $1/(\alpha * \text{sigmoid}(x) + \beta)$, where $\alpha = 8$ and $\beta = 0.1$. As shown in Table 1, the encoder network model's specific structure is part of an end-to-end encoder-decoder architecture.

TABLE 1 The specific structure of the encoder network model.

Name	Input	Kernel size	Stride
conv1	image	7 × 7	1
conv1b	conv1	7 × 7	2
conv2	conv1b	5 × 5	1
conv2b	conv2	5 × 5	2
conv3	conv2b	3 × 3	1
conv3b	conv3	3 × 3	2
conv4	conv3b	3 × 3	1
conv4b	conv4	3 × 3	2
conv5	conv4b	3 × 3	1
conv5b	conv5	3 × 3	2
conv6	conv5b	3 × 3	1
conv6b	conv6	3 × 3	2
conv7	conv6b	3 × 3	1
conv7b	conv7	3 × 3	2

As shown in Table 1, the decoder of the end-to-end network in this paper utilizes deconvolutional operations, taking the output of the second operation in the last layer of the encoder as its input. In the other layers of the decoder, a fusion concatenation operation is employed with the output of the second operation in the second-to-last layer of the encoder. This fusion allows the decoder to access and incorporate more image features from the encoder. The fusion concatenation operation in the other layers of the decoder follows a similar principle. Specifically, each

TABLE 2 Decoder network model specific structure.

Name	Input	Kernel size	Stride
upconv7	conv7b	3 × 3	2
iconv7	[upconv7, conv6b]	3 × 3	1
upconv6	iconv7	3 × 3	2
context	Non-Local Block [upconv6, conv5b]		
iconv6	context	3 × 3	1
upconv5	iconv6	3 × 3	2
iconv5	[upconv5, conv4b]	3 × 3	1
upconv4	iconv5	3 × 3	2
iconv4	[upconv4, conv3b]	3 × 3	1
disp4	iconv4, sigmoid	3 × 3	1
disp4_up	disp4, bilinear	H/4, W/4	
upconv3	iconv4	3 × 3	2
iconv3	[upconv3, conv2b, disp4_up]	3 × 3	1
disp3	iconv3, sigmoid	3 × 3	1
disp3_up	disp3, bilinear	H/2, W/2	
upconv2	iconv3	3 × 3	2
iconv2	[upconv2, conv1b, disp3_up]	3 × 3	1
disp2	iconv2, sigmoid	3 × 3	1
disp2_up	disp2, bilinear	H, W	
upconv1	iconv2, sigmoid	3 × 3	2
iconv1	[upconv1, disp2_up]	3 × 3	1
disp1	iconv1, sigmoid	3 × 3	1
output	[disp1, disp2, disp3, disp4]		

layer in the decoder consists of two operations: deconvolution and concatenation. The deconvolution process upsamples the feature maps from the encoder to obtain higher resolution image features. In the concatenation operation, the upsampled feature maps obtained through deconvolution are combined with the corresponding layer's feature maps from the encoder. By combining the deconvolution and concatenation operations in the decoder, the network can obtain more detailed and contextually rich feature maps. This allows the decoder to generate more accurate and visually appealing image results.

During the deconvolution process, the lack of contextual information may lead to the loss of some fine details in RGB images, thereby affecting the results of image depth estimation. To address this issue, this paper incorporates a Non-local operation attention mechanism, which calculates the similarity of each pixel to weight the context information of each pixel. By doing so, the network can capture and utilize richer contextual information during the deconvolution process, mitigating the loss of fine details and enhancing the accuracy of image depth estimation.

The specific structure of the decoder network model for the end-to-end network is shown in Table 2. The experimental results show that incorporating a Non-local operation attention mechanism between $[upconv6, conv5b]$ yields the best performance. In image depth estimation, to obtain four different scales of depth maps and upsample the first three scales, bilinear interpolation is commonly used. The sampling rates for the first three scales are 1/4, 1/2, and 1, respectively, when performing bilinear interpolation. Finally, by fusing the three scales of depth maps, the network obtains the final set of four different scales of depth maps $[disp1, disp2, disp3, disp4]$.

2.3. Non-local attention mechanism

In computer vision, incorporating attention mechanisms can help models focus on more important areas of an image, thereby reducing the influence of irrelevant background. Non-local operation is a type of attention mechanism that uses global information to capture long-range dependencies between pixels in an image. Compared to local operations, Non-local operations have a broader receptive field and stronger modeling capabilities. The fundamental concept behind non-local operations is to compute the similarity between each pixel and all other pixels in the image. These similarities are then used to adaptively weight the entire image, allowing the model to better understand the global structure of the image.

Figure 7 shows the schematic diagram of the non-local operation module. In this paper, both the Context Aggregation Module and the Transformation Module have incorporated 1 × 1 convolutions, which can reduce the dimensionality of the input feature map without losing information. The Context Aggregation Module is the core component for implementing non-local operations. Its main function is to measure the relationship between two pixels by calculating metrics such as Euclidean distance or cosine similarity. By computing these metrics, the Context Aggregation Module can determine the similarity or dissimilarity between pixels in the input feature map. This allows the module to capture long-range dependencies and establish the global context within the image, enabling the model to understand the relationships between different pixels and extract important contextual information. The Transformation Module is used to convert the input feature map into a new feature map for further processing. The output of the Transformation Module serves as the input to the next layer, enabling communication and integration of data across different layers. 1 × 1 convolutions have two main purposes: first, to reduce dimensions and decrease the number of channels; second, to introduce non-linear elements to enhance the expressive capability of neural networks.

The mathematical definition of the non-local operation is as follows:

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j) g(x_j) \tag{4}$$

In the above Equation: (1) x is feature map; (2) i represents a spatial position of a point on the input x or output y; (3) The response value at position i is represented by y_i; (4) The variable j iterates over the spatial coordinates of all points on the input x or output y; (5) The variable x_j represents the value at position j on the input data; (6) The function $f(x_i, y_j)$ calculates the similarity between position i and position j of the input data; (7) The function $g(x_j)$ calculates a representation of the input data at position j, which can be understood as a weight for the similarity function $f(x_i, y_j)$; and (8) The final response value y of the Non-local operation at position i is obtained by summing the weighted similarities $f(x_i, y_j)g(x_j)$ of each position j relative to the current position i. This sum is then normalized using the normalization factor $C(x)$, which results in the weighted sum of features from all positions being used as the response value at that specific position.

2.4. Deep smoothing loss function

To reduce errors and uncertainties of the results, smooth constraints can be used in depth estimation. Smooth constraints refer to the reduction of noise and discontinuities in the depth map by limiting the differences between the depths of neighboring pixels. This can be accomplished by adding a smoothing term to the loss function of the depth estimation model. To further improve the accuracy and effect of depth estimation, this paper improves the loss function of the model and adopts a depth smoothing loss. The smoothness error of this loss function can be obtained by calculating the gradients of the depth map. To better represent the variations in depth, the gradient computation is performed in the logarithmic domain of the depth map. Based on experience, discontinuous depth values in the depth map are typically found at the edges of the image. Therefore, the edge of the image to be estimated is used as a penalty factor to limit the smoothing loss. The deep smoothing loss constructed in this paper includes the following three aspects:

(1) Smoothing loss based on gradient computation of the depth map. By computing the gradient of the depth map in the logarithmic domain, we can obtain information about depth variations, thereby enhancing the smoothness of the depth map.

$$\partial_x Z_{\log}^{i,j} = Z_{\log}^{i,j} - Z_{\log}^{i+1,j}, i = 0, 1 \cdots W-1; j = 0, 1 \cdots H-1 \quad (5)$$

$$\partial_x Z_{\log}^{i,j} = Z_{\log}^{i,j} - Z_{\log}^{i,j+1}, i = 0, 1 \cdots W-1; j = 0, 1 \cdots H-1 \quad (6)$$

$$\nabla Z_{\log} = |\partial_x Z_{\log}| + |\partial_y Z_{\log}| \quad (7)$$

In the above equation, ∇Z_{log} represents the logarithmic gradient of the depth map, $\partial_x Z_{log}$ denotes the gradient's horizontal component, and $\partial_y Z_{log}$ corresponds to the gradient's vertical component. The indices i and j represent the row and column indices of the depth map, respectively, while W and H represent the width and height of the depth map.

(2) Smoothing Loss based on Edge information. By utilizing the edge information from the input image as a constraint, the depth map can undergo a more accurate smoothing process.

$$\nabla I_{gray} = |\partial_x I_{gray}| + |\partial_y I_{gray}| \quad (8)$$

In the equation, I_{gray} represents the grayscale image obtained from the RGB image, where each pixel value lies in the range of 0 to 255. $\partial_x I_{gray}$ denotes the horizontal gradient, and $\partial_y I_{gray}$ represents the vertical gradient.

(3) Final depth map smoothing loss. As shown in Equation (9):

$$L_{smoth} = \frac{1}{N} \sum_{i,j} (\nabla Z_{\log}^{i,j} \cdot e^{-\nabla I_{gray}^{i,j}}) \quad (9)$$

In the equation, N represents the total number of pixels in the image.

TABLE 3 Experimental environment parameters.

Project	Environment configuration	Version	Quantity
Operating system	Windows10	21H2	-
Deep learning framework	PyTorch	1.12.0	-
GPU	Nvidia	GTX3090	1
Programming languages	Python	3.10	-
Public datasets	Kitti	-	10,000
Self-built dataset	-	-	1,000

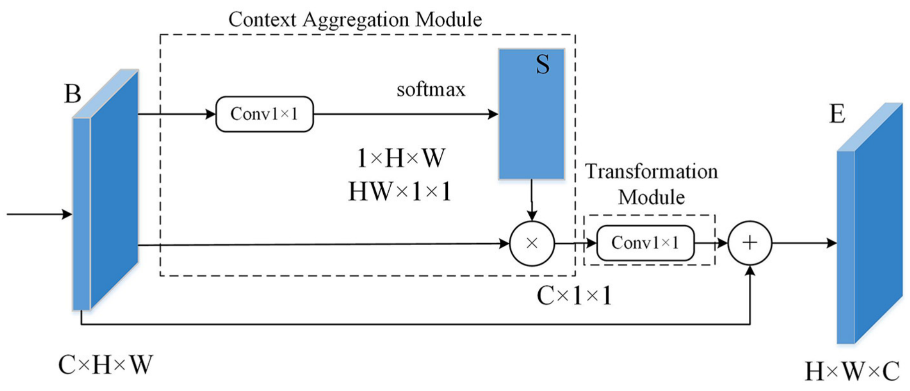


FIGURE 7
Schematic diagram of the non-local operation module.

3. Experiment

3.1. Experimental environment and process

In this study, the improved unsupervised depth estimation network model was implemented and trained using the PyTorch deep learning framework on an NVIDIA GTX3090 GPU. The experiments were conducted to evaluate the model's performance in depth estimation. In addition to using the publicly available KITTI dataset (Geiger et al., 2013), this study also utilized a dataset collected from a catadioptric panoramic camera for the experiments. During the experimental process, batch normalization layers and the Adam optimizer were applied to all layers except the input layer. In the Adam optimizer, set $\beta_1 = 0.95$, $\beta_2 = 0.994$, the learning rate to 0.001, and the mini-batch size to 3. Batch normalization layers are applied to every layer except for the input layer, which helps accelerate the training of the network and improve its accuracy. Moreover, a relatively small mini-batch size was chosen to facilitate faster convergence of the network. Table 3 shows the parameters of the experimental environment in this chapter.

3.2. Evaluation index

This paper employs four evaluation metrics to assess the model's performance, namely Absolute Relative Error (AbsRel), Squared Relative Error (SqRel), Root Mean Squared Error (RMS), and Log Error (Log). The specific form is as follows:

AbsRel: The absolute relative error is a metric used to evaluate the difference between the model's predicted values and the ground truth values. Its calculation formula is the absolute difference between the predicted value and the ground truth value, divided by the ground truth value, reflecting the magnitude of the error relative to the ground truth value.

$$AbsRel = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{|\mathrm{D_i} - \mathrm{D_i^*}|}{\mathrm{D_i^*}} \tag{10}$$

SqRel: The squared relative error is computed by taking the square of the difference between the predicted value and the ground truth value, and then dividing it by the ground truth value.

$$SqRel = \frac{1}{\mathrm{N}} \frac{|\mathrm{D_i} - \mathrm{D_i^*}|^2}{\mathrm{D_i^*}} \tag{11}$$

RMS: The root mean square error is a metric that calculates the square root of the mean of the squared prediction errors. It measures the average magnitude of the prediction errors and is commonly used to evaluate the accuracy of a model's predictions.

$$RMS = \sqrt{\frac{1}{N} \sum_{i=1}^{N} |D_i - D_i^*|^2} \tag{12}$$

TABLE 4 Comparison with other methods.

Methods	AbsRel	SqRel	RMS	Log	Dataset
Eigen	0.204	1.385	5.995	0.283	Kitti
Zhou	0.202	1.347	5.679	0.264	Kitti
In this paper	0.196	1.423	6.237	0.269	Kitti

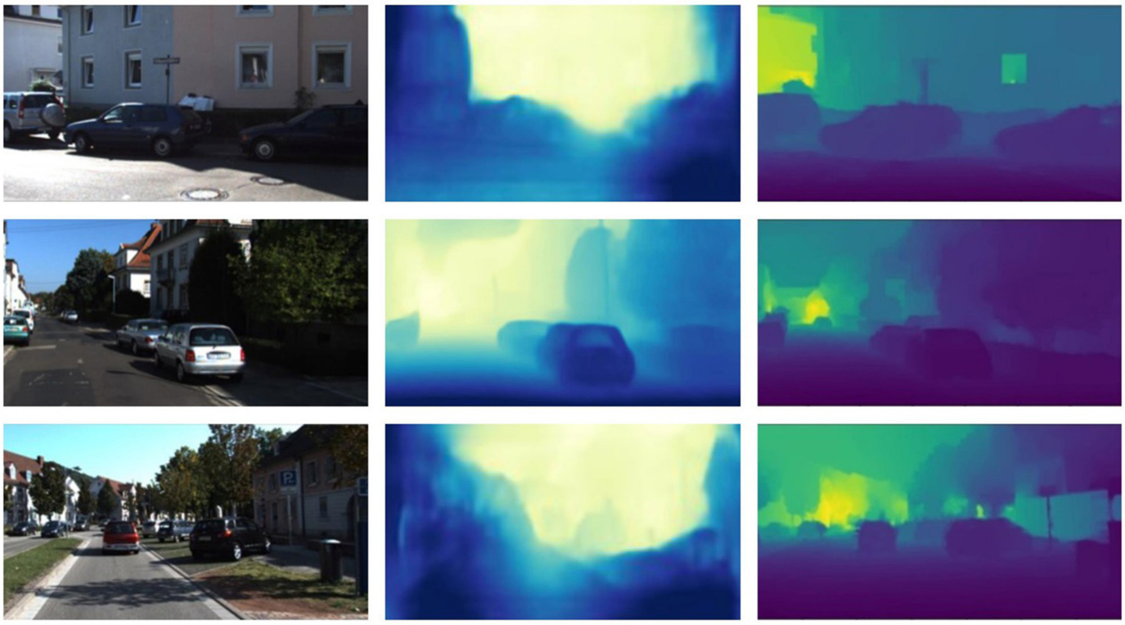


FIGURE 8
Results of depth estimation.

Log: The logarithmic error is a metric that first takes the logarithm of the predicted values and the true values and then calculates the error between these logarithms. This metric is useful when dealing with data that has a large range or significant differences between values.

$$Log = \frac{1}{N}\sum_{i=1}^{N} |\lg D_i - \lg D_i^*| \tag{13}$$

In the above expressions, N represents the total number of valid pixels used for evaluation across all RGB images. D_i denotes the predicted depth of the i-th pixel in the RGB image, and D_i^* represents the true depth of the same pixel.

3.3. Results and analysis

The improved concentric circle approximate expansion algorithm is used to process the panorama and extract the effective area. Creating a dataset from these image segments and performing depth estimation, which validates the robustness of the improved algorithm proposed in this paper.

Figure 8 shows the result of depth estimation. In Figure 8, the first column is the original image, the second column represents the depth estimation results by Zhou et al. (2017), and the last column shows the depth map obtained using the method proposed in this paper. The darker colors in the depth map indicate closer distances, while lighter colors represent farther distances. Through experiments on different scene images, the depth estimation results of the original algorithm are fuzzy, and cannot get accurate results in most cases. The method improved in this paper can generate clearer depth maps. Especially in the case of edge segmentation of objects, the effect of the proposed method is more obvious.

To validate the effectiveness of the improved depth estimation algorithm proposed in this paper, experiments and analyses were conducted on the Kitti dataset. The proposed depth estimation model was evaluated by comparing it with the depth estimation models introduced by Eigen et al. (2014) and Zhou et al. (2017). The experimental results comparison is shown in Table 4.

As shown in Table 4, our proposed method exhibits lower absolute relative error and log error compared to the supervised approach by Eigen et al. (2014), with reductions of 0.8 and 1.4%, respectively. Compared to the unsupervised learning method by Zhou et al. (2017), our approach performs better in terms of absolute relative error, with a reduction of 0.6%, but exhibits slightly higher overall error. In conclusion, our improved method in this paper exhibits better performance in terms of error, with higher accuracy and the ability to address the blurriness issue in image depth estimation.

To further validate the effectiveness of our algorithm, we conducted tests on 200 images captured by the catadioptric panoramic camera in various scenes. Figures 9, 10 show some of the experimental results from different scenes.

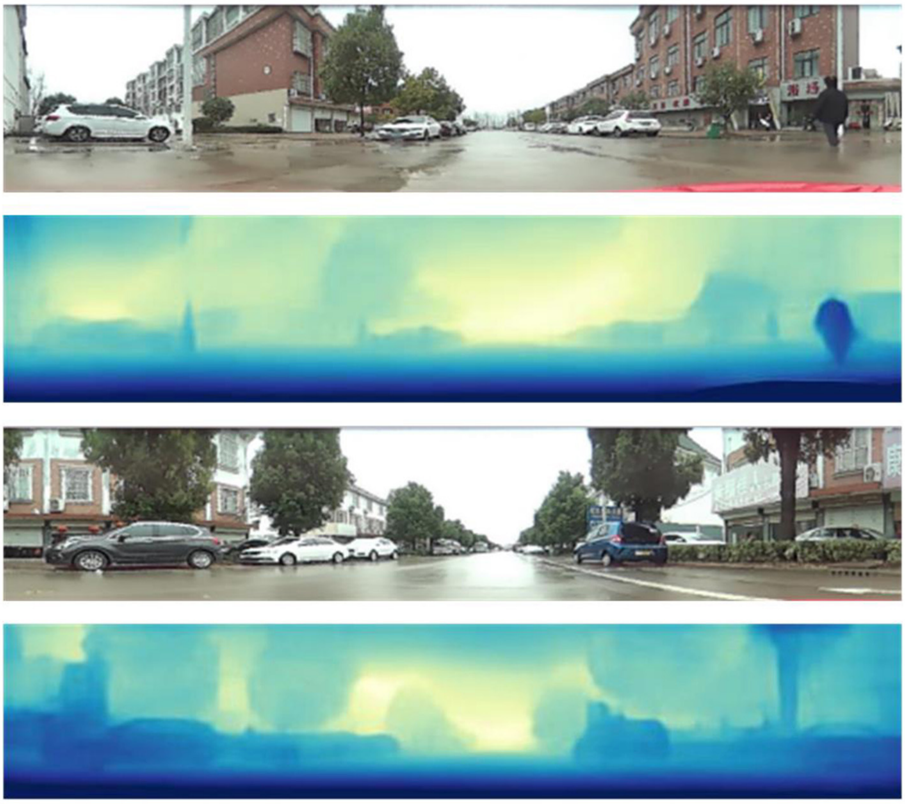


FIGURE 9
Highway depth estimation results.

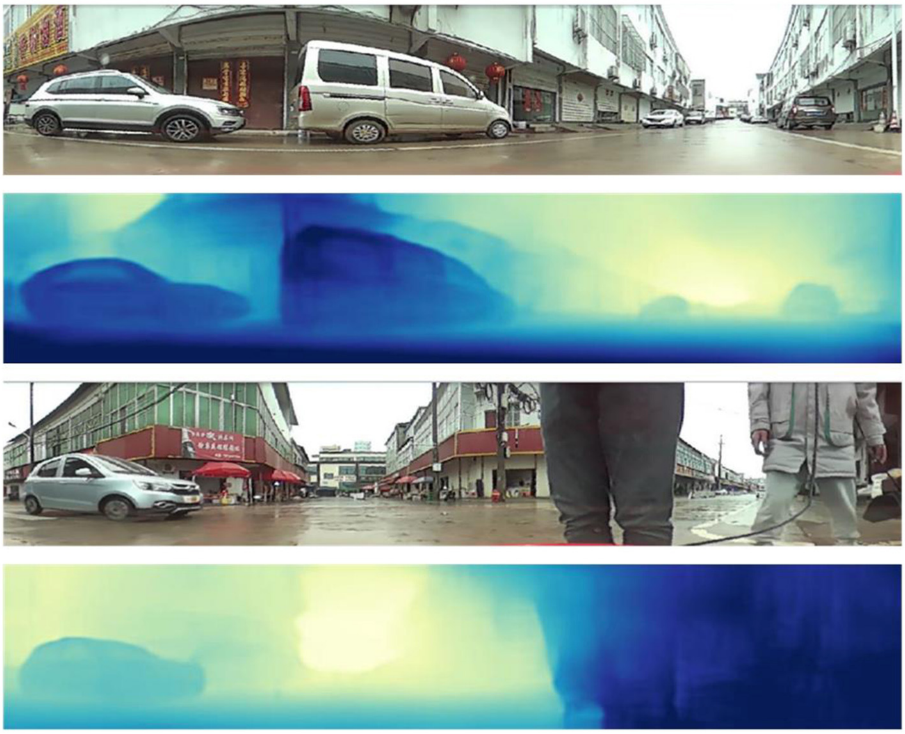


FIGURE 10
Neighborhood street depth estimation results.

In Figures 9, 10, the first and third rows show the original test images. The second and fourth rows display the depth estimation results. The color of the pixels in the depth estimation images represents the distance, where darker colors indicate closer distances and lighter colors indicate farther distances.

From the experimental results, it can be observed that the improved image depth estimation algorithm in this paper can relatively accurately estimate the depth range of objects in the images. Considering the distance analysis relative to the vehicle during image capture, for objects such as vehicles and pedestrians located within a distance of less than 2.5 meters, their corresponding depth values in the depth map fall within the range of 0 to 80, which shows the darkest colors in the depth map; for objects with a distance of 2.5 to 4 meters, the gray values in the depth map results fall within the range of 81 to 150; for objects with a distance greater than 4 meters, the gray values in the depth map results fall within the range of 151 to 255, which results in relatively lighter colors in the depth map.

In conclusion, the research approach proposed in this paper, based on the catadioptric panoramic camera, has demonstrated its effectiveness in depth estimation.

4. Conclusion

This paper proposes a monocular depth estimation algorithm based on the catadioptric panoramic camera. The paper proposes an improved concentric circle approximation unwrapping algorithm to process the panoramic images captured by the catadioptric panoramic camera. This algorithm is used to unwrap the distorted panoramic images into a more usable format for further analysis and depth estimation. The proposed approach enhances the quality and accuracy of the panoramic data. The effective region is extracted according to the unfolded rectangular panorama characteristics. Finally, this paper proposes a new unsupervised end-to-end depth estimation network model. The experimental results show that the depth estimation results of the proposed algorithm are better than the existing algorithms.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

FY: Writing—original draft, Writing—review and editing. LL: Writing—original draft, Writing—review and editing. XD: Writing—original draft. QZ: Writing—review and editing. YL: Writing—review and editing, Writing—original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The Science and Technology Department Project of Jilin Province, Grant No. 20210203039SF.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Badki, A., Troccoli, A., Kim, K., Kautz, J., Sen, P., and Gallo, O. (2020). "Bi3d: Stereo depth estimation via binary classifications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA), 1600–1608. doi: 10.1109/CVPR42600.2020.00167

Baker, S., and Nayar, S. K. (1998). "A theory of catadioptric image formation," in *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)* (Bombay: IEEE), 35–42.

Baker, S., and Nayar, S. K. (1999). A theory of single-viewpoint catadioptric image formation. *Int. J. Comput. Vis.* 35, 175–196. doi: 10.1023/A:1008128724364

Eigen, D., Puhrsch, C., and Fergus, R. (2014). "Depth map prediction from a single image using a multi-scale deep network," in *Advances in Neural Information Processing Systems* 27.

Eldesokey, A., Felsberg, M., Holmquist, K., and Persson, M. (2020). "Uncertainty-aware cnns for depth completion: Uncertainty from beginning to end," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 12014–12023. doi: 10.1109/CVPR42600.2020.01203

Garg, R., Bg, V. K., Carneiro, G., and Reid, I. (2016). "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14 , 2016, Proceedings, Part VIII 14* (Springer), 740–756. doi: 10.1007/978-3-319-46484-8_45

Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *Int. J. Robot. Res.* 32, 1231–1237. doi: 10.1177/0278364913491297

Godard, C., Mac Aodha, O., and Brostow, G. J. (2017). "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 270–279. doi: 10.1109/CVPR.2017.699

Jaramillo, C., Valenti, R. G., and Xiao, J. (2016). "Gums: A generalized unified model for stereo omnidirectional vision (demonstrated via a folded catadioptric system)," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Daejeon: IEEE), 2528–2533. doi: 10.1109/IROS.2016.7759393

Khurana, M., and Armenakis, C. (2018). Localization and mapping using a non-central catadioptric camera system. *ISPRS Ann. Photogram. Rem. Sens. Spat. Inf. Sci.* 4, 145–152. doi: 10.5194/isprs-annals-IV-2-145-2018

Kusupati, U., Cheng, S., Chen, R., and Su, H. (2020). "Normal assisted stereo depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA), 2189–2199. doi: 10.1109/CVPR42600.2020.00226

Li, J., Klein, R., and Yao, A. (2017). "A two-streamed network for estimating fine-scaled depth maps from single rgb images," in *Proceedings of the IEEE International Conference on Computer Vision* 3372–3380. doi: 10.1109/ICCV.2017.365

Liu, M., and Liang, N. (2013). "Detection of moving target using improved optical flow method," in *2013 Fourth World Congress on Software Engineering* (IEEE), 311–315. doi: 10.1109/WCSE.2013.57

Liu, Y., Tian, C., and Huang, Y. (2016). Critical assessment of correction methods for fisheye lens distortion. *Int. Arch. Photogram. Rem. Sens. Spat. Inf. Sci.* 41, 221–228. doi: 10.5194/isprsarchives-XLI-B1-221-2016

Mahjourian, R., Wicke, M., and Angelova, A. (2018). "Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Seattle, WA), 5667–5675. doi: 10.1109/CVPR.2018.00594

Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., et al. (2016). "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 4040–4048. doi: 10.1109/CVPR.2016.438

Nichols, J. M., Waterman, J. R., Menon, R., and Devitt, J. (2010). Modeling and analysis of a high-performance midwave infrared panoramic periscope. *Opt. Eng.* 49, 113202–113202. doi: 10.1117/1.3505866

Svoboda, T., Pajdla, T., and Hlaváč, V. (1998). Epipolar geometry for panoramic cameras," in *Computer Vision–ECCV'98: 5th European Conference on Computer Vision Freiburg, Germany, June, 2–6 , 1998 Proceedings* (Springer), 218–231. doi: 10.1007/BFb0055669

Yamazawa, K., Yagi, Y., and Yachida, M. (1995). "Obstacle detection with omnidirectional image sensor hyperomni vision," in *Proceedings of 1995 IEEE International Conference on Robotics and Automation* (Nagoya: IEEE), 1062–1067. doi: 10.1109/ROBOT.1995.525422

Zhan, H., Garg, R., Weerasekera, C. S., Li, K., Agarwal, H., and Reid, I. (2018). "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 340–349. doi: 10.1109/CVPR.2018.00043

Zhou, T., Brown, M., Snavely, N., and Lowe, D. G. (2017). "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 1851–1858. doi: 10.1109/CVPR.2017.700

TYPE Original Research
PUBLISHED 06 October 2023
DOI 10.3389/fnbot.2023.1280773

Vision-force-fused curriculum learning for robotic contact-rich assembly tasks

Piaopiao Jin[1], Yinjie Lin[2], Yaoxian Song[1], Tiefeng Li[1]* and Wei Yang[1]

[1]Department of Engineering Mechanics, Center for X-Mechanics, Zhejiang University, Hangzhou, China, [2]Hikvision Digital Technology Company, Ltd., Hangzhou, Zhejiang, China

Check for updates

OPEN ACCESS

EDITED BY
Long Jin,
Lanzhou University, China

REVIEWED BY
Jialiang Fan,
University of Nottingham, United Kingdom
Jiahao Chen,
Academy of Sciences (CAS), China
Changkun Du,
Beijing Institute of Technology, China

*CORRESPONDENCE
Tiefeng Li
✉ litiefeng@zju.edu.cn

RECEIVED 21 August 2023
ACCEPTED 19 September 2023
PUBLISHED 06 October 2023

CITATION
Jin P, Lin Y, Song Y, Li T and Yang W (2023) Vision-force-fused curriculum learning for robotic contact-rich assembly tasks. *Front. Neurorobot.* 17:1280773. doi: 10.3389/fnbot.2023.1280773

Contact-rich robotic manipulation tasks such as assembly are widely studied due to their close relevance with social and manufacturing industries. Although the task is highly related to vision and force, current methods lack a unified mechanism to effectively fuse the two sensors. We consider coordinating multimodality from perception to control and propose a vision-force curriculum policy learning scheme to effectively fuse the features and generate policy. Experiments in simulations indicate the priorities of our method, which could insert pegs with 0.1 mm clearance. Furthermore, the system is generalizable to various initial configurations and unseen shapes, and it can be robustly transferred from simulation to reality without fine-tuning, showing the effectiveness and generalization of our proposed method. The experiment videos and code will be available at https://sites.google.com/view/vf-assembly.

KEYWORDS

contact-rich manipulation, multimodal perception, sensor fusion, curriculum learning, robotic assembly task

1. Introduction

In recent years, there has been a growing interest in developing advanced robotic systems capable of performing complex assembly tasks (Sergey et al., 2015; Oikawa et al., 2021; Spector and Zacksenhouse, 2021). These tasks often involve intricate manipulation of objects in contact-rich environments, requiring the robot to possess a high degree of dexterity and adaptability. The success of contact-rich assembly tasks relies on a combination of accurate perception, precise control, and intelligent decision-making. Robots must be equipped with sensory capabilities that enable them to perceive and understand their environment, such as vision systems that capture high-resolution images or depth maps (Morrison et al., 2019; Andrychowicz et al., 2020; Zeng et al., 2021). Additionally, force perception and control mechanisms play a crucial role in managing the physical interaction between the robot and the objects, ensuring gentle and accurate manipulation (Raibert and Craig, 1981; Whitney et al., 1982; Hogan, 1984; Khatib, 1987).

While significant progress has been made in the utilization of unimodal approaches, focusing solely on vision or force (Chhatpar and Branicky, 2001; Tang et al., 2016; Bogunowicz et al., 2020; Stevŝić et al., 2020; Xie et al., 2023), the integration of these modalities presents a compelling opportunity for robots to exploit the complementary nature of vision and force information. By integrating these modalities, robots can enhance their perception and control capabilities, enabling them to adapt effectively to uncertain and dynamic environments. There are two primary approaches to integrating these two modalities: sensor-based controller integration and sensory data fusion (Hosoda et al., 1996). Firstly, visual servoing control and force control are designed separately to form a result scheme capable of coordinating two sensors, and a hybrid structure of sensor-based controllers is built accordingly. Gao and Tedrake (2021) extract the key point representation of the object with a visual detector and then command the robot to the desired pose with the force controller. However, this decoupling method of pose control and force perception ignores the fact that the contact force aroused during the interaction helps to localize the target pose and may enhance the performance of the control scheme. Secondly, given the prioritization of external sensor-based controller coordination over sensory data coordination during the perception phase (Hosoda et al., 1996), this kind of method remains underdeveloped until the emergence of data-driven methodology. This methodology facilitates the fusion of modalities, irrespective of their individual characteristics, and has sparked a surge of interest in numerous studies focusing on robotics perception (Van Hoof et al., 2016; Lee et al., 2020a; Song et al., 2021; Zhao et al., 2021; Spector et al., 2022).

To overcome the limitations of the aforementioned existing methods, we consider a holistic approach to unifying the perception and control modeling process for contact-rich assembly tasks. Specifically, a novel robotic framework based on multimodal fusion and curriculum learning is proposed to improve the performance of contact-rich policy generation end-to-end. Firstly, multimodal perception (i.e., vision and force) are considered to extract multimodal fusion features. Next, we employ reinforcement learning techniques (Sutton and Barto, 2018) to generate both motion and force commands reactive to the multimodal features. For efficient multimodal policy learning, our method includes a two-step vision-force curriculum learning (CL) scheme (Bengio et al., 2009), allowing agents to learn from a curriculum of tasks that progress in complexity and difficulty. The acquired policy is then implemented by a Cartesian motion/force controller, an innovation from our prior work (Lin et al., 2022), designed to guarantee compliant movements amidst uncertain contacts.

To acquire the multimodal policy, we propose a simulated assembly environment based on MuJoCo (Todorov et al., 2012), where the multimodal fusion and policy generation mechanisms are developed. After learning the multimodal policy in simulation, we transfer the simulated system to its physical counterpart. Our multimodal perception-control system could handle the imperfect modeling of interactions in simulated contact-rich scenarios and demonstrate the possibility of a direct sim-to-real transition using a variety of domain randomization techniques (Peng et al., 2018; Chebotar et al., 2019). To evaluate the effectiveness of our proposed framework, a comprehensive series of experiments are conducted on both simulated and physical robots. The results illustrate the remarkable capabilities of the vision-force perception and control system in the simulated environment. It achieves an impressive success rate of 95.3% on a challenging square assembly task whose clearance is 0.1 mm. Furthermore, the algorithm exhibits robust generalization across various spaces, sizes, and even previously unseen shapes. Most notably, the simulated system is seamlessly transferred to the physical environment, achieving zero-shot capabilities and highlighting its potential for real-world implementation.

In summary, the contribution of this work could be summarized as below:

- We propose a novel vision-force framework for contact-rich assembly tasks, enabling multimodal perception and control in challenging and precise operations.
- We introduce a vision-force-fused curriculum learning approach, which progressively coordinates multimodal features based on task difficulty. This innovative approach enables effective vision-force fusion and policy learning specifically tailored to precise assembly tasks.
- We conduct extensive experiments to validate the efficacy of our proposed method. The vision-force perception and control system demonstrates robust generalization capabilities across varying poses and previously unseen shapes. Moreover, we successfully transfer the control scheme to real-world scenarios, ensuring its reliability and applicability in practical settings.

2. Related work

2.1. Force and vision perception in the assembly task

For unimodal perception and control, several methods develop force controllers and map the contact force to misalignment between the peg and the hole (Tang et al., 2016; Inoue et al., 2017). Unten et al. (2023) accurately estimate the relative position between the peg and hole through the force/torque sensing from the transient responses. However, the above methods require prior knowledge of geometry and fail to generalize over new shapes. Apart from the use of force, the utilization of vision to search for holes has also been investigated (Schoettler et al., 2019; Nair et al., 2023). Utilizing an in-hand RGB-D camera, Zhang et al. (2023) develop a 6-DoF robotic assembly system for multiple pegs.

For multimodal perception and control, the complementary nature of vision and force inspires a flurry of study on how to utilize better visual and force sensory feedback. The normal practice is to control the force along the constraint direction while controlling motion via visual servoing along the remaining directions (Haugaard et al., 2021). The task geometry needs to be known a priori in order to properly design the controller through a selection matrix that ensures orthogonality between vision and force control directions. The combination of visual servoing control and impedance control is also actively proposed. The position of the hole is estimated using two depth cameras, followed by a spiral search for the hole using impedance control in Triyonoputro et al. (2019). However, the aforementioned algorithms only combine disparate sensors with their respective controllers. This sensory data separation does not fully exploit the complementarity of vision and force. To better coordinate vision and force, several works have focused on combining visual servoing control and force regulation to achieve a fusion of visual and force perception. The External/Hybrid vision-force control scheme is developed to reach visual and force references simultaneously (Mezouar et al., 2007). The external wrench is transformed into a displacement of the image's feature reference.

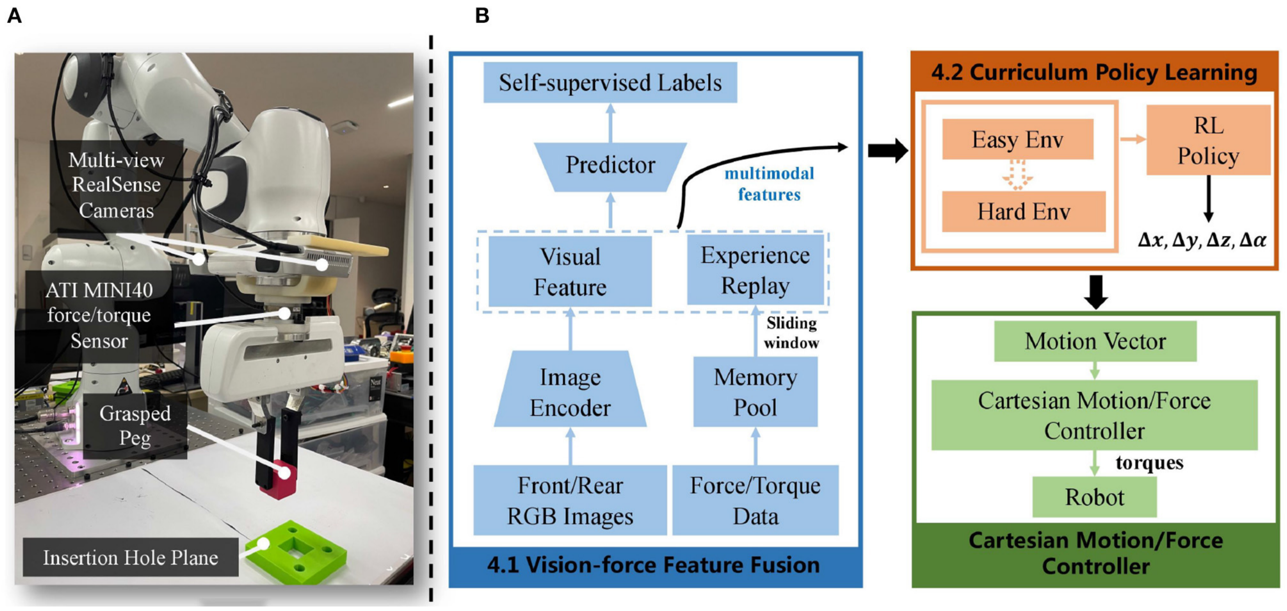


FIGURE 1
(A) Setup of the task: the experimental setup comprises a Franka Emika Panda robot arm equipped with two wrist-mounted RealSense D435 cameras for vision perception and a six-axis ATI mini40 force/torque sensor for interaction forces capturing. **(B)** The overview of our framework includes vision-force feature fusion (blue), followed by curriculum learning-based policy generation (orange), and ended with the motion vector execution module using a Cartesian motion/force controller (green).

And all directions of the task space are simultaneously controlled by both vision and force. Oliva et al. (2021) further generalize the control scheme by not specifying the visual features.

This paper takes a different approach by simultaneously leveraging visual and force features to generate compliant motion and force commands. The system's capability to accommodate environmental variations is greatly expanded as the accurate interaction model is unnecessary in our approach.

2.2. Reinforcement learning-based manipulation

Reinforcement learning (RL) endows robots the promise to accommodate variations in environmental configurations. Some previous works on impedance, admittance, and force control are revisited under the RL scope (Luo et al., 2019; Zang et al., 2023). Oikawa et al. (2021) extend the traditional impedance control using a non-diagonal stiffness matrix learned over RL for precise assembly. Similarly, the use of RL in the admittance control trains the deep neural network that maps task specifications to corresponding parameters (Spector and Zacksenhouse, 2021). Although these algorithms could handle uncertainty and achieve the task, the validness of the unimodal methods is restricted to the single modality's functioning ranges. The development of multimodal policy holds the potential to further enhance manipulation ability (Luo et al., 2021). Lee et al. (2020b) learn a representation model that combines vision, haptics, and proprioceptive data. The state representation is validated in peg-in-hole insertion tasks. Nevertheless, the complicated multimodal features and tedious fine-tuning may hinder practical applications. To simplify the multimodal policy learning process, some strategies leverage prior task knowledge or human demonstrations (Zhao et al., 2021; Spector et al., 2022). Despite their impressive performance in physical insertion experiments, these approaches necessitate human interventions, which are infeasible to acquire in hazardous environments.

Despite the potential of acquiring general policies with RL, the sample inefficiency of RL results in tedious policy training and ill-posed real machine deployment. To overcome the disadvantage, model-based methods (Luo et al., 2019) have been utilized by several researchers to fill this gap, avoiding extensive interactions and training. Curriculum learning (CL) which allows the agents to learn from a curriculum of tasks that progressively increase in complexity and difficulty, could facilitate learning efficiency and improve manipulation performance. Dong et al. (2021) train the insertion agent in progressively more complex environments (wall→corner→U→hole). The result shows that the curriculum training scheme improves the data efficiency of RL and made the problem feasible to solve in a reasonable training time.

In this paper, we propose a novel framework for multimodal curriculum policy learning which could not only explore the compatibility of vision and force but also achieve effective multimodal decision-making. The method is free of human interventions and task priors that expand the scheme's applicability. To effectively deploy the method on the real machine, we train the system in the simulation and then transfer the trained policy to reality. The inconsistencies in perception and control in simulated and real environments (called the reality gap) are bridged by domain randomization (Peng et al., 2018).

3. Problem statement

Our algorithm aims to develop a vision-force perception and control system and validate the scheme in the assembly task. The task is to insert the grasped square peg into the corresponding hole whose clearance is up to 0.1 mm and depth up to 10 mm as shown in Figure 1. Starting from a randomized robot arm configuration, the robot must maneuver and rotate the peg to insert into the target hole, which could be denoted as $robot_{init} \rightarrow hole_{target}$. To reach $hole_{target}$, we formulate the task as a servoing problem and generate the incremental motion vector ΔX at each timestep. The desired robot pose X_{target} could be derived from the current robot pose X_{cur} as:

$$\begin{aligned} X_{target} &= X_{cur} + \Delta X, \\ \Delta X &= f(x_v, x_f), \end{aligned} \tag{1}$$

where x_v and x_f represent raw vision and force observation from robotic sensors, respectively. f is the function mapping from the raw sensory data to the motion vector $\Delta X \in \mathcal{R}^4$ (i.e. $[\Delta x, \Delta y, \Delta z, \Delta\theta]$), where Δx represents the incremental displacement along x-axis, and so does Δy and Δz. $\Delta\theta$ represents incremental z-axis roll command. Absent any prior information about the hole's geometry and pose, the robot must rely solely on sensory feedback to generate motion vector ΔX. Since the robot exhibits distinct dynamic properties before and during contact, some methods split the task into two stages: vision-based hole searching in the free space and force-based insertion in the constraint space. In contrast, our method proposes a single strategy that unifies the two stages, eliminating the need for prior knowledge of how to solve the task and simplifying the modeling process.

Nevertheless, unifying the two stages and devising a single policy function f is quite challenging because visual and force data exhibit different characteristics in the two stages. Therefore, this paper explores the utilization of modality-specific encoders to fuse vision and force and curriculum policy learning to generate motion commands progressively. By leveraging modality-specific encoders, visual and force features are extracted from x_v and x_f, respectively. Through curriculum policy learning, the policy function π_{mlp} automatically generates motion vector ΔX based on the concatenation of visual and force features as shown in Equation (2).

$$\begin{aligned} \phi_v &= E_{vision}(x_v), \\ \phi_f &= E_{force}(x_f), \\ \Delta X &= \pi_{mlp}(\phi_v \oplus \phi_f), \end{aligned} \tag{2}$$

where E_{vision} and E_{force} represent the visual and force encoders, respectively. ϕ_v and ϕ_f the extracted visual and force features, while $(\phi_v \oplus \phi_f)$ concatenation of visual and force features. To this end, the initial servoing problem defined in Equation (1) is transformed into investigating modality-specific encoders and a vision-force-fused curriculum policy learning scheme to generate the incremental motion vector. As such, the target motion vector is derived as in Equation (3). The target motion vector X_{target} is then executed by the Cartesian motion/force controller proposed in our previous work (Lin et al., 2022).

$$X_{target} = X_{cur} + \pi_{mlp}(\phi_v \oplus \phi_f). \tag{3}$$

4. Method

As is shown in our control framework Figure 1, our method begins by using modality-specific encoders to extract visual and force features. These features are then combined to form the multimodal features (Section 4.1). Next, the curriculum policy learning mechanism is employed to train an assembly policy, which hierarchically uses the multimodal features in an environment that gradually increases in difficulty (Section 4.2). Lastly, to execute the motion vector, we utilize the Cartesian motion/force controller proposed in our previous work (Lin et al., 2022). The implementation details are explained in Section 4.3. By coordinating vision and force in the generation and execution of the motions, our vision-force perception and control scheme could fully utilize the multimodality and form a resultant robust assembly system.

4.1. Vision-force feature fusion

The heterogeneous nature of visual and force sensory feedback requires modality-specific encoders to capture the unique characteristics of each modality. We design modality-specific encoders and fusion modules to approximate Equation (2). For the force encoder E_{force}, we employ experience replay with a sliding window of the most recent five frames to extract the force feature. The aggregated force signals are later flattened to a 30-dimensional force feature ϕ_f. Compared to the instant F/T data, the experienced force/torque (F/T) sensory data within the time windows provides a more compact representation of the robot-environment interactions. To further process the data, the raw force data is normalized with the mean (f_μ) and variance (f_{σ^2}). The tanh function further scales the data between -1 and 1.

For the visual encoder E_{vision}, we propose a self-supervised algorithm to extract its RGB feature ϕ_v. As shown in Figure 1, two cameras are symmetrically placed to the gripper. From the top-down view, the grasped peg and hole are observable from the images. With these two images, the visual feature related to the spatial relationship between the peg and hole can be extracted. The spatial relationship between the grasped peg and hole could be denoted by four parameters, E_x, E_y, E_z, and E_θ, which individually represent the translation error along the x, y, and z axes, as well as the z-axis rotational error (Figure 2). To extract the visual feature, the self-supervised neural network predicts three Booleans related to E_x, E_y, and E_θ, while E_z is not observable due to the loss of depth information. Rather than regressing to the values of E_x, E_y, and E_θ, the outputs indicate whether they are positive or negative. More precisely, a label of 0 is assigned when the value is negative, and a label of 1 is assigned when the value is positive.

As illustrated in Figure 3, the first step is to crop two RGB images to a size of 224 × 224. These images are then processed individually using the ResNet50 backbone network (He et al., 2016) and reduced to a 128-dimensional feature space. The

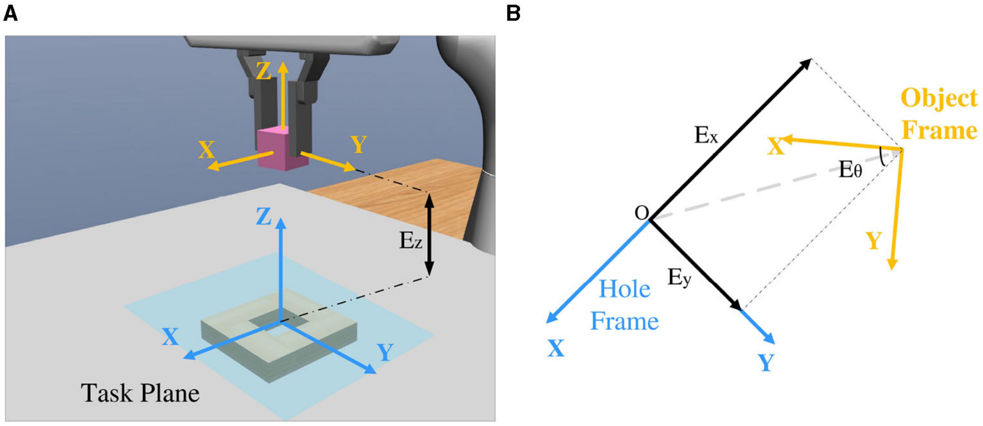


FIGURE 2
(A) Frames of the hole and object in the simulator MuJoCo. **(B)** The transformation between the hole and object frames is denoted by four parameters, E_x, E_y, E_z, and E_θ.

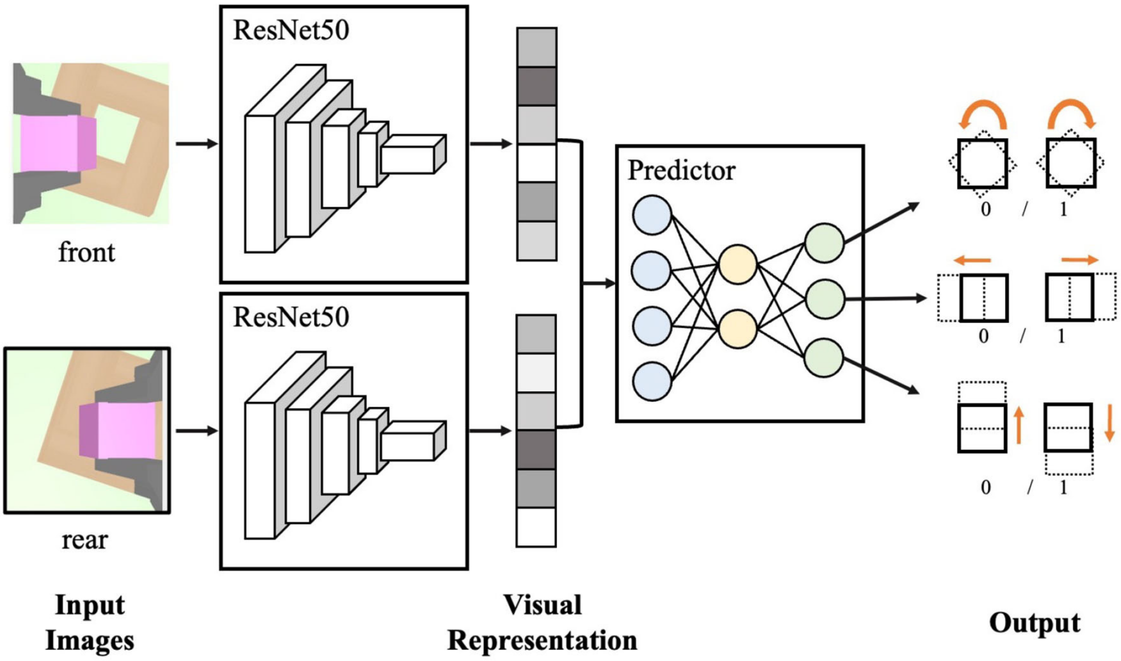


FIGURE 3
The neural network architecture of the self-supervised visual encoder.

resulting visual feature is subsequently input to a three-layer multi-layer-perceptron (MLP) to predict the spatial relationship between the grasped peg and the hole. To train the self-supervised visual neural network, the dataset comprising 60k synthetic multi-view RGB images and labels is collected in the simulation. While this simplifies the labor of performing the operation on real machines, the reality gap of the images hinders the direct transfer of the synthetic visual system to the real robot. To bridge the reality gap, a series of domain randomization techniques are applied, such as Gaussian blurring, white noise, random shadows, and random crops. What's more, in simulation, the colors of the peg, hole, and background are also randomly varied.

4.2. Curriculum policy learning

Our goal is to enable robots to perform precise assembly tasks leveraging visual and force sensory feedback. To achieve the goal, we utilize deep reinforcement learning to map the visual and force sensory data to the robot's motion vector and guide the robot to the target pose following Equation (3). The input to the multimodal policy is the fusion of the visual and force features ($\phi_v \oplus \phi_f$) as defined in Equation (2). π_{mlp} is the multi-layer-perceptron (MLP) function mapping the sensory features to the incremental robot vector ΔX. To learn the policy, the assembly task is formulated as a model-free reinforcement learning problem. This approach avoids the need for an accurate dynamics model that is typically hard to obtain due to the presence of rich contacts. Furthermore, we apply curriculum learning (CL) to structure the task difficulty in accordance with the sensory data input so as to facilitate learning efficiency and enhance model performance. The algorithm is detailed in Algorithm 1.

```
Data: visual feedback $x_v$, force feedback $x_f$,
      and stage $S$
Result: vision-force manipulation policy $\phi_{mlp}$
initiate $S \leftarrow 1$   ▷ train the visual policy in
  stage 1 with 0.5 mm clearance;
if $S = 1$ then
    $\phi_v \leftarrow E_{vision}(x_v)$ ;
    $\Delta X \leftarrow \phi_{init\_mlp}(\phi_v)$   ▷ visual policy $\phi_{init\_mlp}$;
    set the observation and action as $\phi_v$ and $\Delta X$
      and update the PPO policy $\phi_{init\_mlp}$ until it
      converges;
end
initiate $\phi_{mlp}$ with $\phi_{init\_mlp}$, $S \leftarrow 2$   ▷ resume
  vision-force training in stage 2 with 0.1 mm
  clearance;
if $S = 2$ then
    $\phi_v \leftarrow E_{vision}(x_v)$ ;
    $\phi_f \leftarrow E_{force}(x_f)$ ;
    $\Delta X \leftarrow \phi_{mlp}(\phi_v \oplus \phi_f)$ ;
    set the observation and action as $(\phi_v \oplus \phi_f)$
      and $\Delta X$ and update the PPO policy $\phi_{mlp}$ until
      it converges ;
end
```

Algorithm 1. Vision-force-fused curriculum policy learning.

The CL approach divides the training process into two stages: the pure visual policy learning stage and the continued vision-force policy learning stage (shown in Figure 4 and Algorithm 1). The observation space of the first stage contains only 128-dimensional visual feature ϕ_v (Section 4.1), and the larger peg-hole clearance makes this stage of the task easier to manipulate. The difficulty of the second stage intensifies by narrowing the peg-hole clearance to 0.1 mm. We extend the observation space to 158 dimensions by combining the 30-dimensional force feature ϕ_f (Section 4.1). The visual strategy learned in the first stage provides a rough translational and rotational relationship between the grasped peg and the hole. After mastering the required skills in the first stage, the robot proceeds to train in more challenging scenarios incorporating force data. The training in the second stage is like fine-tuning the global visual policy with the local contact force. The action space ΔX for both stages is a 4-dimensional vector representing the desired displacements along x, y, and z axes, and the z-axis rotation roll in the object frame ($\Delta X = [\Delta x, \Delta y, \Delta z, \Delta\theta]$). Meanwhile, to achieve compliance along the z-axis, we command the interaction force along the z-axis to be zero. The Cartesian motion/force controller proposed in Lin et al. (2022) executes the motion and force commands.

Although complex reward functions are often devised for reinforcement learning algorithm (Lee et al., 2020b), sparse rewards are sufficient in our proposed method experimentally. Specifically, the agent obtains the reward of 0.5 if the peg is aligned with the hole and half inserted. The agent gets another reward of 0.5 if the peg is entirely in the hole. Besides, if the peg falls off the gripper, the agent will receive a penalty of −0.2. Since in our setup, the peg is grasped and not fixed to the gripper. The peg can easily fall off the gripper if a large contact force and undesired movements occur.

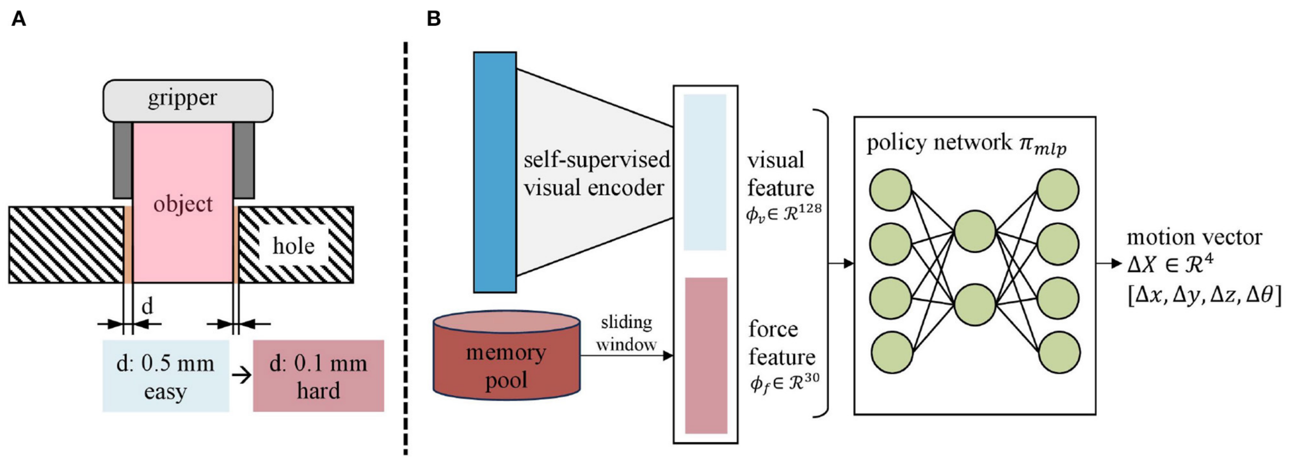

A
B
gripper
object
hole
d
d: 0.5 mm
easy
d: 0.1 mm
hard
self-supervised
visual encoder
memory
pool
sliding
window
visual
feature
$\phi_v \in \mathcal{R}^{128}$
force
feature
$\phi_f \in \mathcal{R}^{30}$
policy network π_{mlp}
motion vector
$\Delta X \in \mathcal{R}^4$
$[\Delta x, \Delta y, \Delta z, \Delta\theta]$

FIGURE 4
The curriculum policy learning procedure. **(A)** The clearance influences policy learning critically. **(B)** Firstly, the peg-hole clearance d is 0.5 mm and the observation is a 128-dimensional visual feature ϕ_v. Secondly, the peg-hole clearance is narrowed to 0.1 mm and the observation space is expanded with the incorporation of force feature ϕ_f. The action space is a four-dimensional motion vector ΔX consisting of the desired displacement along the x, y, and z axes and the z-axis rotation roll.

4.3. Implementation details

To train the self-supervised visual encoder E_{vision} proposed in Section 4.1, we use a binary-class cross-entropy loss to optimize the network with Adam optimizer. We train the network for 20 epochs with batch size 32 and learning rate $1e^{-4}$ under PyTorch 1.11. To achieve a more generalized and robust policy π_{mlp} (Section 4.2), simulation training is conducted under diverse conditions. The initial relative pose of the peg and hole is sampled from a uniform distribution. Specifically, the pose error along the x and y axes is randomly distributed between -10 mm and $+10$ mm, while the z-axis positional error is distributed between 5 mm and 20 mm. The z-axis rotational error is uniformly distributed between -10° and $+10^\circ$. It is assumed that the gripper has already grasped the peg using a human-designed grasp pose. To introduce additional positional randomness, errors along the x and z axes are uniformly distributed between -2 and $+2$ mm. The training of the policy employs Proximal Policy Optimization (PPO) (Schulman et al., 2017), implemented using the stable baselines library (Hill et al., 2018). In training the PPO algorithm, the n_steps is chosen to be 64, and the batch_size is 32, and the gae_lambda to be 0.998.

5. Experiment

We conduct simulated and physical experiments to evaluate the performance and effectiveness of our vision-force perception and control system for the contact-rich assembly task. In particular, we investigate the following four research questions (**RQs**):

- RQ1. How does our proposed method outperform existing work in contact-rich assembly tasks?
- RQ2. Is the multimodal-based policy robust to unseen shapes, colors, and places?
- RQ3. How do modules of our proposed framework improve the final performance?
- RQ4. Can our proposed method perform well in real-world scenarios?

5.1. Evaluation metrics

We define a trial as successful if the robot effectively navigates the peg, securing it within the hole to a depth of 10 mm. Conversely, a trial is considered unsuccessful if the peg slips from the robot's grasp, preventing its insertion into the hole.

5.2. Simulation results analysis

For **RQ1**, we initially evaluate the performance of our vision-force system in the square peg insertion task and then compare the results with those of existing vision-force assembly systems, enabling a comprehensive assessment of the proposed approach. Experimental results indicate that our proposed method outperforms existing baseline work broadly. As shown in Table 1, comparing our method with the baseline from Lee et al. (2020b), we achieve more than 15% improvement in *success rate* (78% → 95.2%). Their method is consistent with ours in fusion vision and force perception and adoption of an impedance controller for incremental motion execution. Nevertheless, they utilize naive RL for policy training while we take a CL approach and split the task into two parts to learn the insertion strategy progressively. Moreover, our Cartesian motion/force controller is more advantageous when dealing with unknown contacts. These two major aspects explain our model's great outperformance. For *clearance*, our method improves 50% relative to baseline from Gao and Tedrake (2021) (0.2 mm → 0.1 mm). Their approach involves a vision-based key point detector followed by a force controller. Our approach differs in formulating the insertion task as a servoing problem and making decisions leveraging both visual and force data end-to-end, thereby achieving more precise manipulation. Although our approach doesn't achieve the high success rate as the work in Spector et al. (2022), our method doesn't require human demonstrations and prior task information. Moreover, our evaluation metrics are stricter by requiring a 10 mm insertion depth while the work in Spector et al. (2022) only requires a 1 mm insertion depth.

For **RQ2**, we first conduct a series of insertion tasks initiating with a randomized peg-hole position error within [-15 mm, 15 mm] along both x and y axes. At each position, we conduct 50 trials to statistically evaluate the system's performance. Next, we test the system's out-of-domain performance on three different shapes that have never been exposed before, namely the pentagonal, triangular, and circular pegs. Experimental results demonstrate that our multimodal system is robust to varying in-domain initial configurations and novel shapes. As shown in Figure 5A, our method achieves an overall success rate of 95.2% across the varying initial pose errors up to 3 cm, which is a reasonable setup in factories and social industries. When the positional error is small than 1.5 cm, the success rate even reaches nearly 100%. The method's robustness to varying positions owns the object-centric design of the observation and action. Specifically, the observation and action are centered on the object coordinate regardless of the robot configurations and global positions. As long as the hole plane can be observable from the in-hand cameras, the robot is able to approach the hole. For novel shapes, the result in Figure 5B indicates the method's remarkable robustness to unseen shapes. Although the novel shapes are never explored before, they share similar task structures with the square pegs. Among the three new shapes, the pentagonal peg is most similar to the square peg and thus has better generalization ability than the other shapes. The triangular peg insertion task is more challenging with a higher z-axis roll requirement. Surprisingly, the model behaves poorly on the circular peg, probably due to the small contact surface (line contact) between the peg and the gripper. Although the hardware setup for the circular peg easily causes slippage and tilt, it still maintains a success rate of 60%.

5.3. Ablation study of proposed module

For **RQ3**, we investigate the contributions of the design choices, namely the act of vision-force perception fusion and the curriculum vision-force fusion mechanism.

This section conducts two comparisons: (1) we compare whether the fusion of vision and force boost performance over vision only. (2) we investigate whether the two-stage curriculum learning (CL) fusion mechanism could improve fusion efficiency and manipulation performance than the naive reinforcement learning (RL) fusion mechanism. To verify the suppositions mentioned above, we design the following models:

TABLE 1 The performance of different multimodal models in the assembly task.

Models	Clearance ↓	Peg	Modalities	DoF	Success rate ↑	Shape generalization	Human demonstration
Gao and Tedrake (2021)	0.2 mm	Unfixed	RGB/depth/force	3	74%	No	No
Lee et al. (2020b)	2 mm	Fixed	RGB/depth/force	4	78%	Yes	No
Spector et al. (2022)	–	Unfixed	RGB/force	6	**97.5%**	No	Yes
Ours	**0.1 mm**	Unfixed	RGB/force	4	95.2%	Yes	No

The bold values represent the best performance among the comparisons.

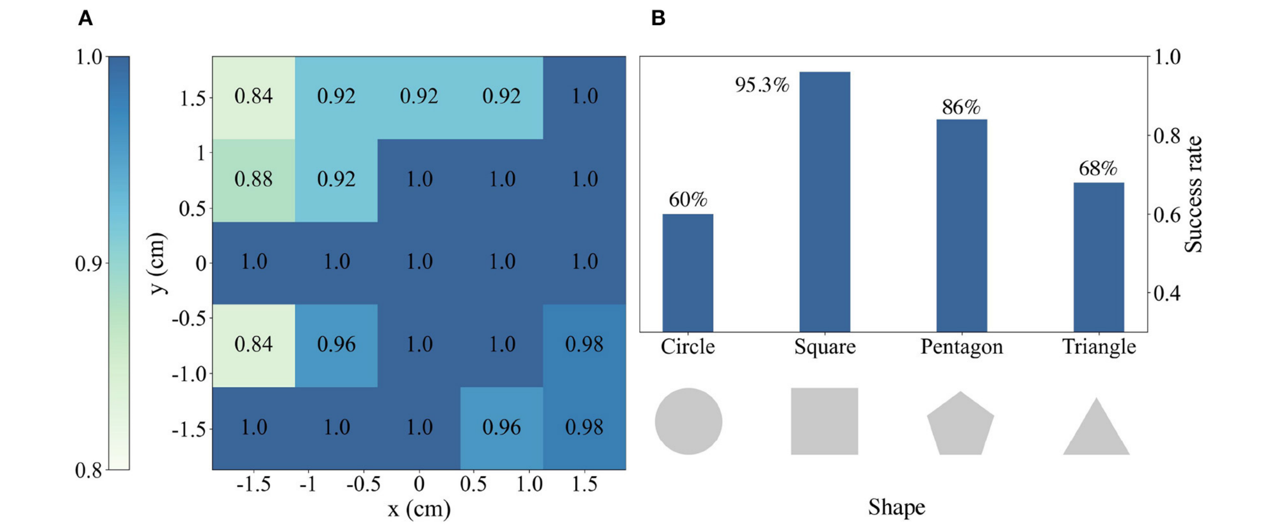


FIGURE 5
(A) Simulation experimental results with varied initial positions for peg-hole operations using a square object. Each individual value corresponds to the insertion success rate at that region, thereby providing a comprehensive overview of the spatial distribution and variations in success rates of the square peg insertion task. **(B)** The success rate of different peg-hole objects, in which square is used in training (in-domain) while others only are used to test (out-of-domain).

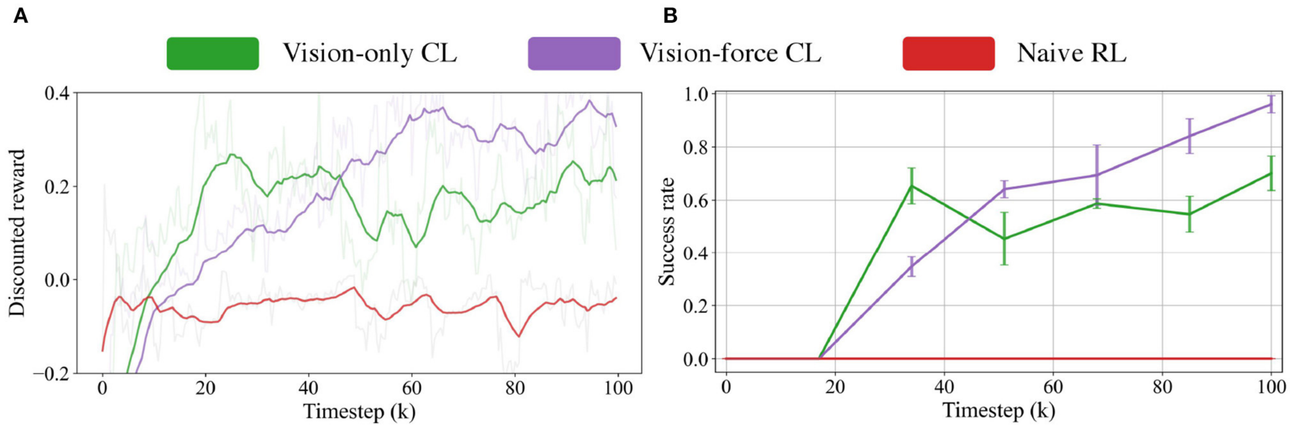


FIGURE 6
(A) Training curves of three models, including the Vision-only CL model, Vision-force CL model, and Naive RL model. **(B)** The insertion success rates at different training stages of three models.

- **Vision-only CL model** contains only vision perceptually and curriculum learns the visual policy.
- **Vision-force CL model** curriculum learns the vision-force multimodal policy.
- **Naive RL model** naively learns the vision-force policy with RL.

All of the above-mentioned models are trained and tested in simulation. For a fair comparison, all the models except the *Naive RL model* are initialized using a pure visual policy trained with a larger clearance. Figures 6A, B visualize the learning curves during the training and the test results for 250 trials with three random seeds.

5.3.1. Vision-force vs. vision-only

The experiment results indicate the superior performance of the *Vision-force CL model* over the *Vision-only CL model*, manifesting the necessity of vision-force fusion in contact-rich precise manipulation tasks. As demonstrated in Figure 6, comparing the *Vision-force CL model* with the *Vision-only CL model*, the proposed method achieves more than 20% improvement in success rate (70% → 95.2%). Although the ablative *Vision-only CL model* doesn't perform as well as *Vision-force CL model*, it maintains a success rate of 70% which indicates that integrating sensor-based controllers is a solution for contact-rich tasks. Formulating the assembly task as a servoing problem and solving it with curriculum policy learning end-to-end is a good fit for the challenging precise insertion. Nonetheless, the fusion of vision and force perception results in significantly improved outcomes, as the contact-rich insertion task is sensitive to both visual and force signals. Vision perception serves as the main data stream to locate the target, and force perception is a complementary data source when contacts are made and interactions occur.

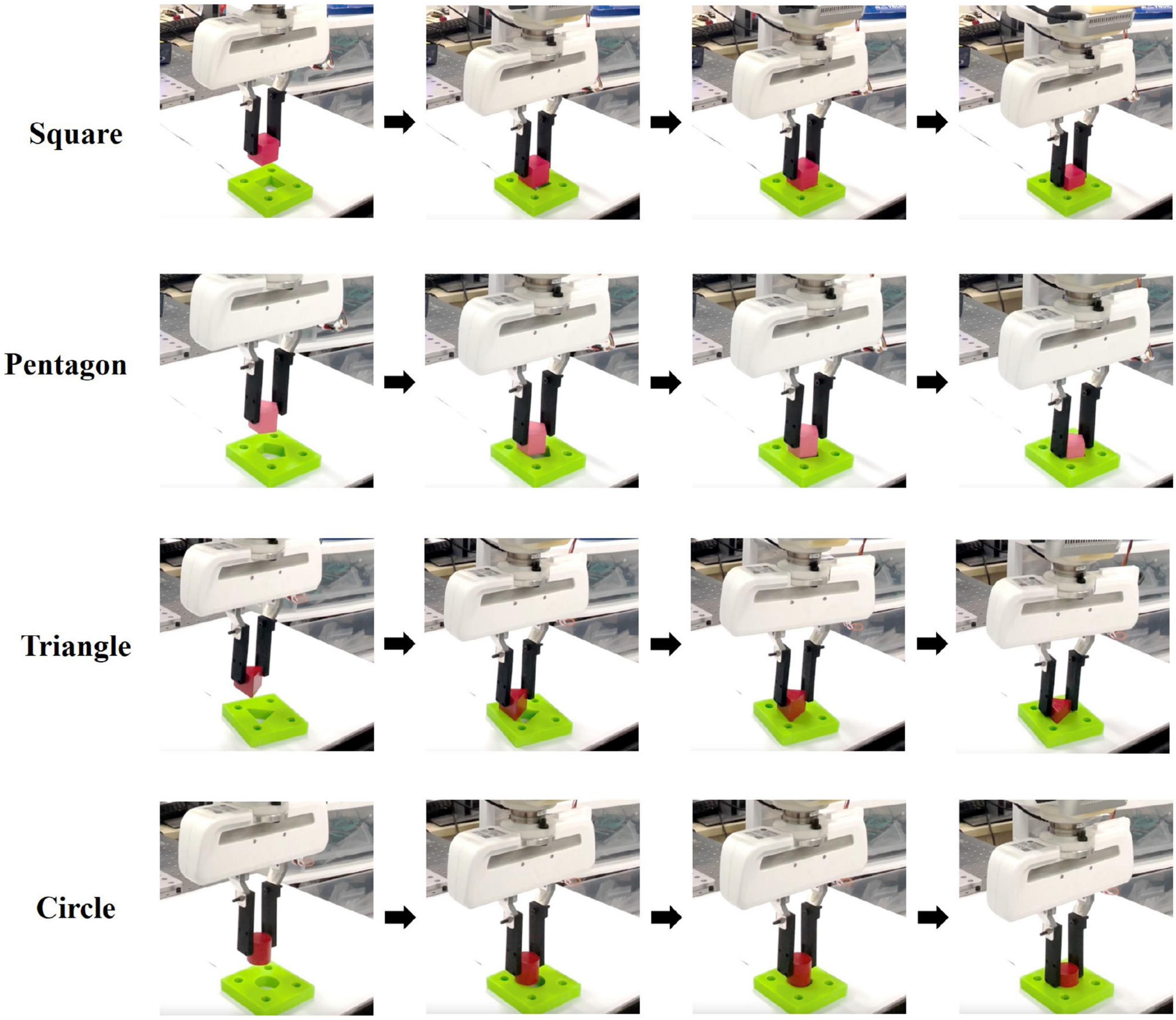


FIGURE 7
Snapshots of the peg-hole insertion process during the physical robot experiments.

TABLE 2 Performance on physical assembly task.

Models \ Shapes	Square	Pentagon	Triangle	Circle
Vision-only CL	3/10	8/10	3/10	2/10
Vision-force CL	6/10	9/10	5/10	4/10

5.3.2. CL-based model vs. naive RL model

In terms of the results of the CL, the experiment results indicate that the conduct of CL is decisive for multimodal strategy generation in extremely challenging tasks. Comparing the *Vision-force CL model* with the *Naive RL model* in Figure 6, the proposed method could achieve a remarkable success rate of 95.2%. In contrast, the ablative *Naive RL model* couldn't succeed in the task and has 0% success rate. The huge performance gap between the two models comes from the different policy learning formulations. The *Naive RL model* leverages visual and force data to insert the square peg whose clearance is as low as 0.1 mm from scratch. Nevertheless, it's difficult for the agent to coordinate the motions and insert the peg into the hole as a rash motion will cause the slippery of the peg and finally lead to the local optima of the algorithm. Different from the naive RL modeling, the CL-based modeling first learns a visual policy on a larger clearance and is followed by the fusion of force perception on a 0.1 mm clearance task. The curriculum task difficulty organization provides a more effective policy generation approach.

5.4. Physical robot experiments

For **RQ4**, we perform direct sim-to-real transfer and generalization tests on the real machine. In the experiment, the robot first grasps the object and then executes the assembly policy to insert the peg into the hole. The insertion hole is rigidly fixed so as not to add extra compliance to the system. Figure 7 shows the four shapes utilized in our experiments, along with snapshots captured during the insertion process. Specifically, the square, pentagonal, triangular, and round peg-hole clearances are 0.37 mm, 0.44 mm, 1 mm, and 0.41 mm, respectively. Table 2 presents the results obtained from the experiments on these four shapes using two models: the *Vision-only CL model* and the *Vision-force CL model*. Experiment results indicate that the simulated assembly system can be transferred to the physical robot. Moreover, the *Vision-force CL model* demonstrates stronger robustness against the ablative *Vision-only CL model*. As shown in Table 2, the *Vision-force CL model* achieves 20% success rate more than the *Vision-only CL model*. Although the *Vision-only CL model* could be transferred to the physical robot, the *Vision-force CL model* even demonstrates better behavior. The performance gap between the two models is consistent with that in the simulated system. Although dynamics in the simulated and physical environment differ, the domain randomization techniques applied to the visual encoder and the compliant motion/force controller to handle uncertain contacts minimize the reality gap. Furthermore, consistent with the situation in simulations, the method could also be generalized to unseen shapes in physical environments.

6. Conclusion

This paper proposes a novel vision-force fusion scheme for contact-rich precise assembly tasks. Our approach utilizes a curriculum policy learning mechanism to effectively fuse multi-view visual and force features and implement compliant motions. By effectively fusing visual and force data from perception to control, our method achieves higher precision and better generalization to unseen shapes in the simulated environment. The experiments on the physical environment validate the practicability of our simulated system. Our vision-force system significantly contributes to the advancement of multimodal contact-rich tasks.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

PJ: Investigation, Methodology, Writing—original draft, Writing—review and editing. YL: Methodology, Writing—review and editing. YS: Writing—review and editing. TL: Funding acquisition, Writing—review and editing. WY: Writing—review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the following programs: National Natural Science Foundation of China (T2125009 and 92048302). The funding of Laoshan laboratory (Grant no. LSKJ202205300). The funding of Pioneer R&D Program of Zhejiang (Grant no. 2023C03007).

Acknowledgments

The authors would like to thank and appreciate the support they received from the Center for X-Mechanic laboratory for providing the necessary facilities to accomplish this work.

Conflict of interest

YL was employed by Hikvision Digital Technology Company, Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Andrychowicz, O. M., Baker, B., Chociej, M., Jozefowicz, R., McGrew, B., Pachocki, J., et al. (2020). Learning dexterous in-hand manipulation. *Int. J. Rob. Res.* 39, 3–20. doi: 10.1177/0278364919887447

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). "Curriculum learning," in *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09* (Montreal, QC: ACM Press), 1–8. doi: 10.1145/1553374.1553380

Bogunowicz, D., Rybnikov, A., Vendidandi, K., and Chervinskii, F. (2020). Sim2real for peg-hole insertion with eye-in-hand camera. *arXiv*. [preprint]. doi: 10.48550/arXiv.2005.14401

Chebotar, Y., Handa, A., Makoviychuk, V., Macklin, M., Issac, J., Ratliff, N., et al. (2019). "Closing the sim-to-real loop: adapting simulation randomization with real world experience," in *Proc. IEEE Int. Conf. Robot. Automat.* (Montreal, QC: IEEE), 8973–8979. doi: 10.1109/ICRA.2019.8793789

Chhatpar, S., and Branicky, M. (2001). "Search strategies for peg-in-hole assemblies with position uncertainty," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Vol. 3 (Maui, HI: IEEE), 1465–1470. doi: 10.1109/IROS.2001.977187

Dong, S., Jha, D. K., Romeres, D., Kim, S., Nikovski, D., Rodriguez, A., et al. (2021). "Tactile-rl for insertion: generalization to objects of unknown geometry," in *Proc. IEEE Int. Conf. Robot. Automat.* (Xi'an: IEEE), 6437–6443. doi: 10.1109/ICRA48506.2021.9561646

Gao, W., and Tedrake, R. (2021). KPAM 2.0: feedback control for category-level robotic manipulation. *IEEE Robot. Autom. Lett.* 6, 2962–2969. doi: 10.1109/LRA.2021.3062315

Haugaard, R., Langaa, J., Sloth, C., and Buch, A. (2021). "Fast robust peg-in-hole insertion with continuous visual servoing," in *Proceedings of the 2020 Conference on Robot* Learning, Volume 155 of Proceedings of Machine Learning Research, 1696–1705. PMLR.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. doi: 10.1109/CVPR.2016.90

Hill, A., Raffin, A., Ernestus, M., Gleave, A., Traore, R., Dhariwal, P., et al. (2018). *Stable Baselines*. San Francisco, CA: GitHub Repository.

Hogan, N. (1984). "Impedance control: an approach to manipulation," in *1984 American Control Conference* (San Diego, CA: IEEE), 304–313. doi: 10.23919/ACC.1984.4788393

Hosoda, K., Igarashi, K., and Asada, M. (1996). "Adaptive hybrid visual servoing/force control in unknown environment," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, *Volume* 3 (Osaka: IEEE), 1097–1103. doi: 10.1109/IROS.1996.568956

Inoue, T., De Magistris, G., Munawar, A., Yokoya, T., and Tachibana, R. (2017). "Deep reinforcement learning for high precision assembly tasks," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Vancouver, BC: IEEE), 819–825. doi: 10.1109/IROS.2017.8202244

Khatib, O. (1987). A unified approach for motion and force control of robot manipulators: the operational space formulation. *IEEE J. Robot. Autom.* 3, 43–53. doi: 10.1109/JRA.1987.1087068

Lee, M. A., Yi, B. Martín-Martín, R., Savarese, S., and Bohg, J. (2020a). "Multimodal sensor fusion with differentiable filters," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* (Las Vegas, NV: IEEE), 10444–10451. doi: 10.1109/IROS45743.2020.9341579

Lee, M. A., Zhu, Y., Zachares, P., Tan, M., Srinivasan, K., Savarese, S., et al. (2020b). Making sense of vision and touch: learning multimodal representations for contact-rich tasks. *IEEE Trans. Robot.* 36, 582–596. doi: 10.1109/TRO.2019.2959445

Lin, Y., Chen, Z., and Yao, B. (2022). Unified method for task-space motion/force/impedance control of manipulator with unknown contact reaction strategy. *IEEE Robot. Autom. Lett.* 7, 1478–1485. doi: 10.1109/LRA.2021.3139675

Luo, J., Solowjow, E., Wen, C., Ojea, J. A., Agogino, A. M., Tamar, A., et al. (2019). "Reinforcement learning on variable impedance controller for high-precision robotic assembly," in *Proc. IEEE Int. Conf. Robot. Automat.* (Montreal, QC: IEEE), 3080–3087. doi: 10.1109/ICRA.2019.8793506

Luo, J., Sushkov, O., Pevceviciute, R., Lian, W., Su, C., Vecerik, M., et al. (2021). Robust multi-modal policies for industrial assembly via reinforcement learning and demonstrations: a large-scale study. *arXiv*. [preprint]. doi: 10.48550/arXiv.2103.11512

Mezouar, Y., Prats, M., and Martinet, P. (2007). "External hybrid vision/force control," in *Proc. Int. Conf. Adv. Robot.* (Jeju), 170–175.

Morrison, D., Corke, P., and Leitner, J. (2019). "Multi-view picking: next-best-view reaching for improved grasping in clutter," in *Proc. IEEE Int. Conf. Robot. Automat.* (Montreal, QC: IEEE), 8762–8768. doi: 10.1109/ICRA.2019.8793805

Nair, A., Zhu, B., Narayanan, G., Solowjow, E., and Levine, S. (2023). "Learning on the job: self-rewarding offline-to-online finetuning for industrial insertion of novel connectors from vision," in *2023 IEEE International Conference on Robotics and Automation (ICRA)* (London: IEEE), 7154–7161. doi: 10.1109/ICRA48891.2023.10161491

Oikawa, M., Kusakabe, T., Kutsuzawa, K., Sakaino, S., and Tsuji, T. (2021). Reinforcement learning for robotic assembly using non-diagonal stiffness matrix. *IEEE Robot. Autom. Lett.* 6, 2737–2744. doi: 10.1109/LRA.2021.3060389

Oliva, A. A., Giordano, P. R., and Chaumette, F. (2021). A general visual-impedance framework for effectively combining vision and force sensing in feature space. *IEEE Robot. Autom. Lett.* 6, 4441–4448. doi: 10.1109/LRA.2021.3068911

Peng, X. B., Andrychowicz, M., Zaremba, W., and Abbeel, P. (2018). "Sim-to-real transfer of robotic control with dynamics randomization," in *Proc. IEEE Int. Conf. Robot. Automat.* (Brisbane, QLD: IEEE), 1–8. doi: 10.1109/ICRA.2018.8460528

Raibert, M. H., and Craig, J. J. (1981). Hybrid position/force control of manipulators. *J. Dyn. Syst. Meas. Control* 103, 126–133. doi: 10.1115/1.3139652

Schoettler, G., Nair, A., Luo, J., Bahl, S., Ojea, J. A., Solowjow, E., et al. (2019). "Deep reinforcement learning for industrial insertion tasks with visual inputs and natural rewards," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Las Vegas, NV: IEEE), 5548–5555. doi: 10.1109/IROS45743.2020.9341714

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv*. [preprint]. doi: 10.48550/arXiv.1707.06347

Sergey, L., Wagener, N., and Abbeel, P. (2015). "Learning contact-rich manipulation skills with guided policy search," in *Proc. IEEE Int. Conf. Robot. Automat.*, 156–163.

Song, Y., Luo, Y., and Yu, C. (2021). Tactile–visual fusion based robotic grasp detection method with a reproducible sensor. *Int. J. Comput. Intell. Syst.* 14, 1753–1762. doi: 10.2991/ijcis.d.210531.001

Spector, O., Tchuiev, V., and Di Castro, D. (2022). "Insertionnet 2.0: minimal contact multi-step insertion using multimodal multiview sensory input," in *2022 International Conference on Robotics and Automation (ICRA)* (Philadelphia, PA: IEEE), 6330–6336. doi: 10.1109/ICRA46639.2022.9811798

Spector, O., and Zacksenhouse, M. (2021). "Learning contact-rich assembly skills using residual admittance policy," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* (Prague: IEEE), 6023–6030. doi: 10.1109/IROS51168.2021.9636547

[Stevšić, S., Christen, S., and Hilliges, O. (2020). Learning to assemble: estimating 6D poses for robotic object-object manipulation. *IEEE Robot. Autom. Lett.* 5, 1159–1166. doi: 10.1109/LRA.2020.2967325

Sutton, R. S., and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT press.

Tang, T., Lin, H.-C., Zhao, Y., Chen, W., and Tomizuka, M. (2016). "Autonomous alignment of peg and hole by force/torque measurement for robotic assembly," in *Proc. IEEE Int. Conf. Autom. Sci. Eng.* (Fort Worth, TX: IEEE), 162–167. doi: 10.1109/COASE.2016.7743375

Todorov, E., Erez, T., and Tassa, Y. (2012). "MuJoCo: a physics engine for model-based control," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* (Vilamoura-Algarve: IEEE), 5026–5033. doi: 10.1109/IROS.2012.6386109

Triyonoputro, J. C., Wan, W., and Harada, K. (2019). "Quickly inserting pegs into uncertain holes using multi-view images and deep network trained on synthetic data," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* (Macau: IEEE), 5792–5799. doi: 10.1109/IROS40897.2019.8968072

Unten, H., Sakaino, S., and Tsuji, T. (2023). Peg-in-hole using transient information of force response. *IEEE/ASME Trans. Mechatron.* 28, 1674–1682. doi: 10.1109/TMECH.2022.3224907

Van Hoof, H., Chen, N., Karl, M., van der Smagt, P., and Peters, J. (2016). "Stable reinforcement learning with autoencoders for tactile and visual data," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* (Daejeon: IEEE), 3928–3934. doi: 10.1109/IROS.2016.7759578

Whitney, D. E. (1982). Quasi-static assembly of compliantly supported rigid parts. *J. Dyn. Syst. Meas. Control* 104, 65–77. doi: 10.1115/1.3149634

Xie, Z., Jin, L., Luo, X., Zhou, M., and Zheng, Y. (2023). A bi-objective scheme for kinematic control of mobile robotic arms with manipulability optimization. *IEEE/ASME Trans. Mechatron.*

Zang, Y., Wang, P., Zha, F., Guo, W., Ruan, S., Sun, L., et al. (2023). Geometric-feature representation based pre-training method for reinforcement learning of peg-in-hole tasks. *IEEE Robot. Autom. Lett.* 8, 3478–3485. doi: 10.1109/LRA.2023.3261759

Zeng, A., Florence, P., Tompson, J., Welker, S., Chien, J., Attarian, M., et al. (2021). "Transporter networks Rearranging the visual world for robotic manipulation," in *Conference on Robot Learning (PMLR)* (Auckland, NZ), 726–747.

Zhang, K., Wang, C., Chen, H., Pan, J., Wang, M. Y., Zhang, W., et al. (2023). "Vision-based six-dimensional peg-in-hole for practical connector insertion," in *2023 IEEE International Conference on Robotics and Automation (ICRA)* (London: IEEE), 1771–1777. doi: 10.1109/ICRA48891.2023.10161116

Zhao, T. Z., Luo, J., Sushkov, O., Pevceviciute, R., Heess, N., Scholz, J., et al. (2021). Offline meta-reinforcement learning for industrial insertion. *arXiv*. [preprint]. doi: 10.48550/arXiv.2110.04276

Frontiers in Neurorobotics

TYPE Original Research
PUBLISHED 11 October 2023
DOI 10.3389/fnbot.2023.1285831

OPEN ACCESS

EDITED BY
Long Jin,
Lanzhou University, China

REVIEWED BY
Fei Yan,
Changchun University of Science and Technology, China
Shuaishi Liu,
Changchun University of Technology, China
Hua Wang,
Peking University, China

*CORRESPONDENCE
Peiyuan Jiang
✉ darcy_j@163.com

RECEIVED 30 August 2023
ACCEPTED 25 September 2023
PUBLISHED 11 October 2023

CITATION
Pan W, Jiang P, Li Y, Wang Z and Huang J (2023) Research on automatic pilot repetition generation method based on deep reinforcement learning.
Front. Neurorobot. 17:1285831.
doi: 10.3389/fnbot.2023.1285831

Research on automatic pilot repetition generation method based on deep reinforcement learning

Weijun Pan[1], Peiyuan Jiang[1]*, Yukun Li[1], Zhuang Wang[1] and Junxiang Huang[2]

[1]Air Traffic Control Automation Laboratory, College of Air Traffic Management, Civil Aviation Flight University of China, Deyang, China, [2]Department of Safety Management, Xiamen Air Traffic Management Station, East China Air Traffic Management Bureau, Xiamen, China

Using computers to replace pilot seats in air traffic control (ATC) simulators is an effective way to improve controller training efficiency and reduce training costs. To achieve this, we propose a deep reinforcement learning model, RoBERTa-RL (RoBERTa with Reinforcement Learning), for generating pilot repetitions. RoBERTa-RL is based on the pre-trained language model RoBERTa and is optimized through transfer learning and reinforcement learning. Transfer learning is used to address the issue of scarce data in the ATC domain, while reinforcement learning algorithms are employed to optimize the RoBERTa model and overcome the limitations in model generalization caused by transfer learning. We selected a real-world area control dataset as the target task training and testing dataset, and a tower control dataset generated based on civil aviation radio land-air communication rules as the test dataset for evaluating model generalization. In terms of the ROUGE evaluation metrics, RoBERTa-RL achieved significant results on the area control dataset with ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.9962, 0.992, and 0.996, respectively. On the tower control dataset, the scores were 0.982, 0.954, and 0.982, respectively. To overcome the limitations of ROUGE in this field, we conducted a detailed evaluation of the proposed model architecture using keyword-based evaluation criteria for the generated repetition instructions. This evaluation criterion calculates various keyword-based metrics based on the segmented results of the repetition instruction text. In the keyword-based evaluation criteria, the constructed model achieved an overall accuracy of 98.8% on the area control dataset and 81.8% on the tower control dataset. In terms of generalization, RoBERTa-RL improved accuracy by 56% compared to the model before improvement and achieved a 47.5% improvement compared to various comparative models. These results indicate that employing reinforcement learning strategies to enhance deep learning algorithms can effectively mitigate the issue of poor generalization in text generation tasks, and this approach holds promise for future application in other related domains.

KEYWORDS

controller training, transfer learning, text generation, reinforcement learning, generalization

1. Introduction

In recent research projects (Holone and Nguyen, 2015) and as indicated by the International Civil Aviation Organization (ICAO), it is projected that air traffic flow will continue to grow at an annual rate of 3 to 6% after 2025. Consequently, the demand for Air Traffic Controllers (ATCOs) will increase year by year. ATCOs communicate control instructions to pilots via Very High-Frequency (VHF) radio to manage air traffic. According to safety and reliability regulations in Air Traffic Control (ATC), pilots are required to promptly and accurately repeat control instructions they receive to ensure the correct understanding of instructions issued by ATCOs (Lin et al., 2019). ATCOs undergo specific training, including foundational courses and simulator training, to qualify for working in actual ATC scenarios. Control training simulators typically consist of two seats: one for the controller and the other for the pilot. Completing controller training requires dedicated personnel to control the pilot seat for the repetition and response to control instructions, incurring additional training costs, including equipment and personnel expenses, as illustrated in Figure 1 (Zhang et al., 2022a). In recent years, artificial intelligence (AI) technologies have been widely applied in the ATC domain (Lin, 2015; Srinivasamurthy et al., 2017; Yang et al., 2019). To alleviate the workload of ATCOs, the European Union (EU) has introduced Automatic Speech Recognition (ASR) technology into ATC to reduce their workload (Helmke et al., 2016) and enhance work efficiency (Helmke et al., 2017). Projects funded by Horizon 2020 have also constructed ATCO decision support systems using AI technology to alleviate the workload of ATCOs (Kleinert et al., 2017). These research endeavors aim to assist controllers with intelligent systems to reduce error rates and alleviate workload. Furthermore, enhancing the quality of ATCO training is another approach to reducing potential human errors (Yiu et al., 2021). Some scholars have explored the use of intelligent systems to improve the training efficiency and professionalism of ATCOs, fundamentally reducing human errors. For example, Hoekstra and Ellerbroek (2016) developed an ATC simulator called "BlueSky," which significantly advanced research in air traffic management (ATM) despite its lower level of intelligence. Lin et al. (2021) proposed an AI-based pilot framework for ATCO training, capable of replacing the pilot seat with relatively high confidence. This framework covers several core technologies, including speech recognition, Controlling Instruction Understanding (CIU), Information Extraction (IE), Pilot Repetition Generation (PRG), Text-to-Speech (TTS), and human-computer interaction technology, as illustrated in Figure 2. Zuluaga-Gomez et al. integrated various state-of-the-art AI-based tools to build an automatic captain system, expediting the training process for air traffic controllers (ATCo) (Zuluaga-Gomez et al., 2023). However, the above research primarily focuses on the entire pilot system, with limited in-depth research on the PRG module. Building upon the aforementioned research efforts, this paper delves deeper into the task of PRG and presents novel advancements.

In Figure 1, Area Control Centers (ACC) are responsible for managing the airspace within a designated region, coordinating aircraft flights, and ensuring the orderly flow of air traffic and the tower primarily oversees the Terminal Control Area (TMA), which encompasses the airspace including airports and their surrounding regions. Due to the differences in the scope of controlled airspace, there are significant variations in the content of control instructions, leading to disparities in the data distributions between the two.

The focus of this study is on the PRG, which belongs to the field of Natural Language Processing (NLP) and falls under the task of Natural Language Generation (NLG). We achieved PRG by fine-tuning pre-trained language models based on Transformer and Seq2Seq architectures. Furthermore, we employed the policy gradient algorithm from reinforcement learning to further optimize the model and overcome the issue of poor generalization in transfer learning. The innovations of this paper are as follows: (1) Addressing the characteristics of pilot repetition generation tasks, we transformed the human-machine dialogue problem into a text summarization problem, providing a new perspective for related research. (2) By utilizing transfer learning strategies, we overcame the limitations of insufficient training data in this field, caused by the difficulty of data collection. (3) We used

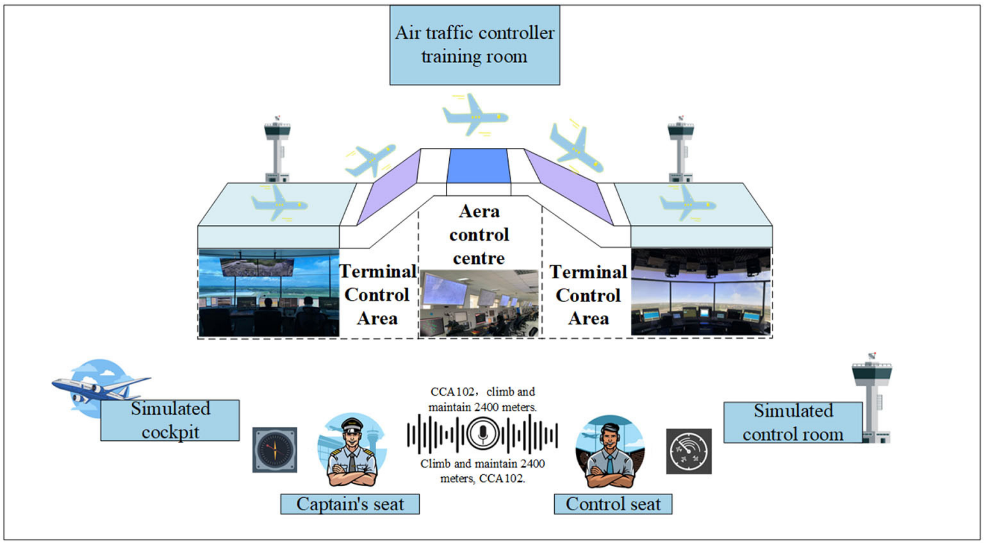


FIGURE 1
ATCOs training process.

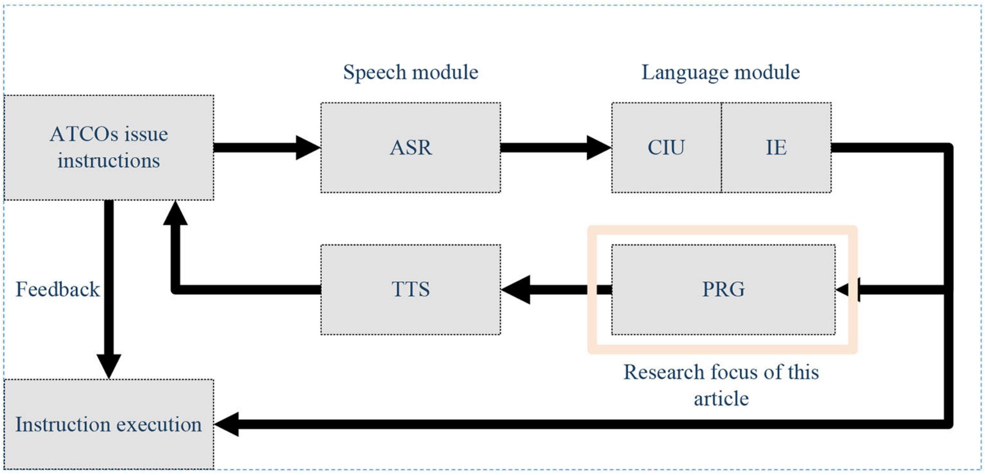


FIGURE 2
Core technologies of automatic pilot seat.

the policy gradient algorithm to optimize the cross-entropy loss function, overcoming the exposure bias issue associated with using cross-entropy loss in text generation tasks and enhancing the generalization of the transfer learning model. (4) We constructed a control instruction text dictionary based on the structural features of control instruction texts. This dictionary enables fine-grained tokenization of control instruction texts, facilitating subsequent metric evaluations. In addition, based on control instruction tokenization, we introduced a keyword-based evaluation to assess the quality of generated pilot repetitions. The introduced keyword evaluation metrics provide an intuitive reflection of the model's performance.

2. Related work

The general characteristics of PRG are as follows: (1) The length of the repetition instructions is generally shorter than that of the control instructions, and for mandatory control instructions, the repetition instructions should be consistent with the meaning of the control instructions. (2) There are fewer instances of ongoing dialogues (similar to single-turn dialogues in human-machine conversations). Based on these characteristics, PRG can be transformed from a human-machine dialogue task to a text summarization task for processing. Currently, text summarization techniques can be classified into extractive summarization and abstractive summarization based on the summarization method (Nazari and Mahdavi, 2019). Extractive summarization extracts keywords based on their importance and forms a summary. However, it only considers the word frequency and does not take into account the semantic information of sentences, resulting in poor coherence of the generated sentences. On the other hand, abstractive summarization summarizes the essential information of sentences through paraphrasing and synonym replacement. Compared to extractive summarization, abstractive summarization has better representation ability and can understand the contextual semantics of sentences. In the task of automatic text summarization, since both the input and output are text sequences, the model needs to pay more attention to the relationship between the semantic information of generated sentences and the coherence of sentences (Liu et al., 2021).

Over the years, the development of automatic text summarization has been slow due to the limitations of statistical-based methods in text representation, understanding, and generation capabilities (Zhang et al., 2019). Recently, with the continuous improvement of neural network theory and technology, deep learning has emerged as one of the most promising approaches and has achieved state-of-the-art results in many tasks (de Souza et al., 2018; Luo et al., 2019; Mane et al., 2020; Miao et al., 2020). Among them, the introduction of automatic text summarization models based on the encoder-decoder architecture has brought new advancements to deep learning-based automatic text summarization (Zhang et al., 2022b). In the current context, with the advancement of sequence-to-sequence frameworks, generative models tend to outperform extractive models (Alexandr et al., 2021).

Most of the research on generative summarization focuses on the encoder-decoder structure of sequence-to-sequence models, addressing various issues in the summarization process by incorporating attention mechanisms, pointer-generator mechanisms, coverage mechanisms, or replacing recurrent neural networks (RNNs) with convolutional neural networks. Rush et al. (2015) were the first to use attention mechanisms on the seq2seq model to address headline generation. To further improve model performance, Nallapati et al. proposed the pointer generator model (Nallapati et al., 2016b), which successfully handles out-of-vocabulary (OOV) words due to limited vocabulary. This model was later improved with the use of coverage mechanisms (See et al., 2017). Since the encoder and decoder in the Seq2Seq architecture are implemented using convolutional neural networks or RNNs, their feature extraction capabilities are not as powerful as the Transformer model. The emergence of the Transformer model based on self-attention architecture has ushered in a new era in NLP, ensuring that models can learn deeper language logic and semantic information of words. Examples of such models include BERT (Devlin et al., 2018), GPT-2 (Radford et al., 2019), Bart, and Roberta. BERT predicts words based on their contextual information, while GPT-2 predicts words based on the preceding context. Therefore, BERT is suitable for natural language understanding (NLU) tasks, while GPT-2 is more suitable for NLG tasks. Inspired by BERT and GPT-2, the Bart model combines the strengths of both, making it more suitable for text generation scenarios compared to BERT and achieving better results than GPT-2 (Lewis et al., 2019). The RoBERTa model (Liu et al., 2019), compared to BERT, GPT-2, and Bart, has advantages in terms of pre-training methods, deeper network structure, larger batch size, and unmasked training, especially for text summarization tasks. These advantages enable RoBERTa to better understand semantics, capture language features, and generate more accurate and coherent text summaries. The proposed deep reinforcement learning model in this paper is based on RoBERTa.

3. Challenges in PRG and our work

3.1. Challenges in PRG

(1) With the increase in the number of parameters in deep learning models, training high-performance models in supervised learning requires a large amount of data. In the field of ATC, data acquisition is extremely challenging due to the confidentiality of the data. Additionally, the obtained raw ATC voice data needs to be professionally annotated, which incurs high annotation costs. These factors pose significant challenges to the application and development of deep learning techniques in this domain. (2) Current NLG models often suffer from poor generalization, and this issue becomes more pronounced in the case of small datasets. Improving model generalization is a challenging task that requires extensive research. (3) Since control instructions are composed of a series of keywords (Pan et al., 2023), evaluating the generated pilot repetition instructions using ROUGE-N and ROUGE-L standards requires the segmentation of the control instructions. This necessitates the construction of a dictionary, adding extra workload. Furthermore, the specific nature of pilot repetition instructions limits the effectiveness of using ROUGE-N and ROUGE-L for evaluating the quality of generated instructions. Therefore, a new evaluation metric is needed to assess the quality of generated pilot repetition instructions.

3.2. Our work

We have conducted in-depth research on text generation. We found that NLG involves three major tasks: neural machine translation (NMT), text summarization, and dialogue response generation (Nallapati et al., 2016a). These tasks share the common characteristic of having text sequences as inputs and outputs, but they also have differences. The difference between text summarization and machine translation lies in the fact that generated summaries are typically very short and not influenced by the length of the source text, while the generated summary and the source text need to be semantically consistent (Zhou, 2012). Furthermore, text summarization involves compressing the source text in a lossy manner while retaining key information, which contradicts the lossless requirement of machine translation (Hastie, 2012). The difference between dialogue response generation and text summarization is that the generated text in dialogue response has logical coherence with its preceding and following context. Currently, there is no unified evaluation criterion for the quality of dialogue generation results (Song et al., 2019). PRG is a special NLG task that belongs to both dialogue response generation and text generation tasks. For certain inquiry instructions (such as “please respond when received"), the nature of their repetition belongs to dialogue, with logical relationships between the preceding and following text. However, most control instructions are mandatory instructions, and the nature of their repetition belongs to text summarization, where the meaning should remain consistent throughout.

Based on the analysis of PRG tasks mentioned above, we have adopted the following strategies from the perspective of text summarization to address the challenges faced by repetition generation. For challenge one, we use transfer learning by pretraining the model on other domain data and fine-tuning it on the target domain to achieve the generation of repetition instructions. For challenge two, we employ the policy gradient algorithm from reinforcement learning to optimize the cross-entropy loss in the pre-trained model. The cross-entropy loss relies on target labels in the training data for parameter optimization. This leads to a significant decrease in model performance when applying the fine-tuned model to similar datasets due to differences in the training label distribution. The core of the policy gradient algorithm is to optimize the parameters of the policy network by evaluating the quality of generated summaries. This allows the model to learn how to generate high-quality summaries rather than generating text summaries similar to the training sample labels, greatly improving the generalization performance of the transfer learning model. Additionally, we compare the effects of fine-tuning current mainstream pre-trained models to demonstrate the effectiveness of our proposed model. For challenge three, to enable a detailed evaluation of model performance and facilitate model improvement, we use a new evaluation criterion to assess the quality of generated repetition instructions. This criterion provides a more accurate reflection of the model’s performance compared to the ROUGE evaluation criterion. Furthermore, we construct a control instruction text dictionary based on the control instruction text dataset. Using the Jieba word segmentation tool, we split the generated instruction text based on coarse-grained and fine-grained information, allowing the calculation of various metrics using computer programs.

4. Methodology

4.1. Proposed framework

Deep Reinforcement Learning (DRL) is a method that combines deep learning and reinforcement learning to solve decision-making problems with high-dimensional state and action spaces. It uses deep neural networks (DNNs) as function approximators to learn value functions or policy functions, enabling end-to-end learning from raw input to action selection. In text summarization tasks, DRL can be used to train models to generate high-quality summaries (Keneshloo et al., 2019; Sun et al., 2021). The application of DRL in text summarization generally follows the basic framework of reinforcement learning. In this framework, an agent learns the optimal policy by interacting with the environment. In this case, the environment consists of the original text and the generated summary, and the agent observes the current text state and selects actions to generate the next word. The reward function provides rewards to the agent based on the quality evaluation of the generated text, with higher rewards indicating higher-quality summaries. The key to applying DRL in text summarization lies in designing appropriate state representations, action spaces, reward functions, and policy networks. State representation refers to transforming the original text into continuous vector representations using word embeddings or encoder networks to capture the semantic and contextual information of the text. The action space defines the operations that the agent can choose, typically selecting the next word to generate from a vocabulary. The reward function is used to evaluate the quality of the generated summary. Language model-based metrics such as ROUGE evaluation can be used as the reward function to measure the similarity between the generated summary and the reference summary. The policy network is a DNN that selects actions to generate the next word based on the current state. RNNs or attention mechanisms can be used to capture the context of the text and make sequential word decisions. By applying DRL to text summarization, the model can learn to generate high-quality summaries through interactions with the environment. During the training process, the agent optimizes the parameters of the policy network to maximize the cumulative reward while generating summaries. This approach allows for end-to-end training on large-scale datasets without the need for manual annotations, leveraging deep learning techniques to extract features from raw input and generate more accurate and fluent summaries.

In our proposed RoBERTa-RL model, we use Word Piece embedding as the state representation of the environment. We use ROUGE-1 as the reward function and RoBERTa as the policy network. The action generation policy is implemented using Beam Search, and parameter updates are performed using the policy gradient algorithm. The architecture of our proposed deep reinforcement learning model, RoBERTa-RL, is illustrated in Figure 3.

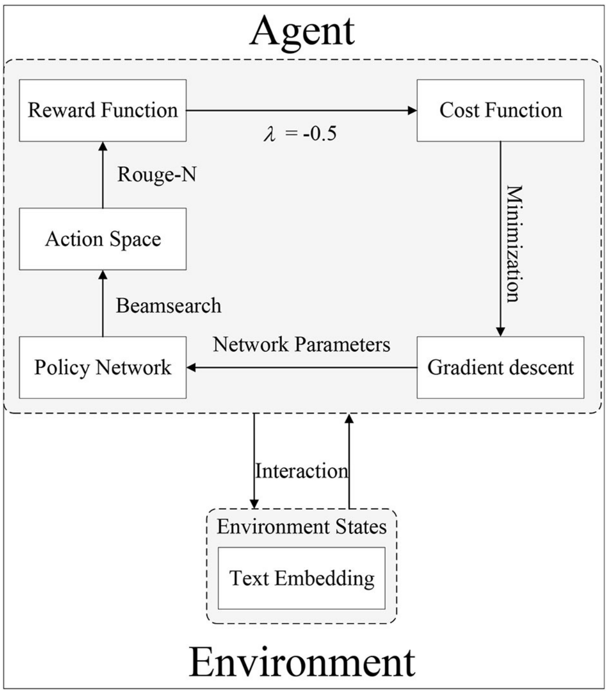


FIGURE 3
The proposed DRL architecture.

4.2. Training process of RoBERTa-RL

Figure 3 provides a detailed description of the training process of the proposed DRL model architecture. Let's assume $S = \{x_1, x_2, ..., x_n\}$ represents the original input text, where $x_1, x_2, ..., x_n$ are input characters. Firstly, S undergoes RoBERTa encoding to convert it into the state representation of the environment, denoted as h_t. This process is described by Equation (1), where $RoBERTa_{\text{embedding}}()$ represents the encoding function:

$$h_t = RoBERTa_{\text{embedding}}(S) \quad (1)$$

The policy network generates the output text y_t based on the state representation h_t of the input environment and the action policy Beam search. The specific process is described by Equation (2), where $Beamsearch()$ represents the action policy function:

$$y_t = Beamsearch\ (RoBERTa\ , h_t) \quad (2)$$

The ROUGE function calculates the reward value R_t based on the generated text y_t and the reference summary $T_{\text{reference}}$. The specific formula is described by Equation (3), where $ROUGE - 1()$ represents the reward function.

$$R_t = ROUGE - 1\left(y_t, T_{\text{reference}}\right) \quad (3)$$

The cost function $COST$ is composed of the weighted sum of the negative average reward value and the cross-entropy loss, where λ is the weight. The specific formula is described by Equation (4).

$$COST = -\lambda\ mean\ (R_t) + (1 - \lambda)\ CrossEntorpyLoss\ \left(y_t, T_{\text{reference}}\right) \quad (4)$$

The policy update is performed using the policy gradient algorithm, which updates the policy network parameters θ based on the gradient of the cost function. The specific formula is described by Equation (5), where α represents the learning rate.

$$\theta = \theta - \alpha \nabla \theta \quad (5)$$

4.3. Evaluation criteria

ROUGE (recall-oriented understudy for gisting evaluation) measures the quality of summaries by calculating the overlap

TABLE 1 Calculation results of ROUGE-1, ROUGE-2, and ROUGE-L for the example.

Evaluation metrics	Number of *n*-grams in the reference instruction	Number of overlapping *n*-grams between the repetition and the reference	Result
ROUGE-1	8	4	0.5
ROUGE-2	7	3	0.429
ROUGE-L	8	4	0.5

units (such as n-grams, word sequences, and word pairs) between the generated summary and the reference summary (Lin and Och, 2004; Elmadani et al., 2020). This evaluation criterion has been widely used for evaluating automatic summarization tasks. ROUGE-1 and ROUGE-2 are used to assess informativeness, while ROUGE-L is used to assess fluency. N is typically set to 1 or 2. The ROUGE-1 and ROUGE-2 scores have been shown to be the most consistent with human judgments. The calculation method for ROUGE-N is described by Equation (6).

$$ROUGE - N = \frac{\sum\limits_{S\in Ref}\sum\limits_{gram_n\in S} Count_{match}(gram_n)}{\sum\limits_{S\in Ref}\sum\limits_{gram_n\in S} \text{Count}(gram_n)} \tag{6}$$

In Equation (6), n represents the length of n-grams, Ref is the set of reference summaries. $Count_{\text{match}}(gram_n)$ is the maximum number of n-grams that appear simultaneously in the generated summary and the corresponding reference summary, while $Count(gram_n)$ is the number of n-grams in the reference summary. The calculation formula for ROUGE-L is described by Equations (7–9).

$$R_{\text{LCS}} = \frac{LCS(C,S)}{len(S)} \tag{7}$$

$$P_{\text{LCS}} = \frac{LCS(C,S)}{len(C)} \tag{8}$$

$$F_{\text{LCS}} = \frac{(1+\beta^2)R_{\text{LCS}}P_{\text{LCS}}}{R_{\text{LCS}}+\beta^2 P_{\text{LCS}}} \tag{9}$$

In Equations (7-9), R_{LCS} represents recall, P_{LCS} represents precision, and F_{LCS} denotes the ROUGE-L value. β is a tunable parameter, and in this paper, it is set to 0.5, indicating that F_{LCS} gives equal importance to R_{LCS} and P_{LCS}.

Due to the specificity of the ATC domain, repetition must be completely accurate to be considered a valid repetition instruction. Pilot repetition instructions require responding to the control instructions based on ATC rules without losing any crucial information. According to ATC rules (Drayton and Coxhead, 2023), ATCO instructions must start with the aircraft identification (ACID) to specify the communicating aircraft, while pilot repetitions should end with their ACID to differentiate them from ATCO instructions. Based on the characteristics of the generated repetitions mentioned above, using only the ROUGE evaluation metric cannot comprehensively assess the model's performance. For example, in the control instruction dataset, the controller issues the Chinese control instruction "MU5424, yi jing xiang Beijing shen qing, xian zan shi bao chi 7500", and the reference repetition instruction is "Yi jing xiang Beijing shen qing, xian zan shi bao chi 7500, MU5424". After word segmentation, the tokens are as follows: "Yi jing/xiang/Beijing/shen qing/zan shi/bao chi/7500/MU5424". When the model generates the result "Zan shi/bao chi/7500/MU5424", evaluating the result using the ROUGE-N and ROUGE-L evaluation methods yields the results shown in Table 1. However, from the perspective of repetition generation rules, this repetition instruction is correct.

From the results in Table 1, it can be seen that although the ROUGE metrics can to a large extent reflect the quality of the generated repetition instructions, there are times when unreasonable situations may arise. Therefore, considering the characteristics of ATC instructions and the repetition criteria, we introduce a new evaluation metric specific to this domain, based on keyword evaluation. The evaluation metrics include Call Sign Accuracy (CSA), Action Instruction Accuracy (AIA), and Parameter Accuracy (PA). Finally, the Total Accuracy (TA) is calculated. Only when an instruction has all three sub-factors correctly, it can be considered as a correct repetition instruction. The definitions and calculation formulas of the specific metrics are as follows: (1) Call sign is composed of the airline abbreviation and flight number, and its accuracy is calculated using the following formula.

$$CSA = \frac{1}{N}\sum_{i=1}^{N} g(i) \tag{10}$$

(2) Action instruction refers to the actions contained in the ATC instruction, such as climb, descend, maintain, etc., and its accuracy is calculated using the following formula.

$$AIA = \frac{1}{N}\sum_{i=1}^{N} q(i) \tag{11}$$

(3) Parameter refers to the key supplementary information of the instruction actions in the ATC instruction, including speed, altitude, heading, waypoints, etc., and its accuracy is calculated using the following formula.

$$PA = \frac{1}{N}\sum_{i=1}^{N} h(i) \tag{12}$$

In Equations (10–12), N represents the number of samples to be tested, and $g(i)$, $q(i)$, and $h(i)$ represent the feature functions of call sign, action instruction, and parameter of the instruction, respectively. The specific formulas is described by Equation (13).

$$g(i), q(i), h(i) = \begin{cases} 1 & if\ pred_i = truth_i \\ 0 & otherwise \end{cases} \tag{13}$$

(4) TA represents the total accuracy, which is the sentence-level accuracy. A generated repetition is considered valid and correct

TABLE 2 Examples of word entries in the dictionary.

Category	Example
Airline abbreviations	Air China, Eastern, CA, MU, Sichuan, 3U, etc.
Numbers	0 ("dong"), 1 ("yao"), 2 ("liang"), 7 ("guai"), etc.
Altitude	600, 900, 1,200, 1,500, . . ., 13,700
Speed	250 knots, 180 knots, etc.
Heading	Direct flight, offset, flying heading, etc.
Waypoint	Dawangzhuang, BUBDA, ANDIN, P23, etc.
Proper noun	Indicated airspeed, field pressure, planned route, instrument flight, etc.

only when the call sign, parameters, and action instructions in the repetition match the ground truth. The specific formulas are described by Equations (14, 15).

$$T(i) = \begin{cases} 1 & if\ g(i) = q(i) = h(i) \\ 0 & otherwise \end{cases} \tag{14}$$

$$TA = \frac{1}{N}\sum_{i=1}^{N} T(i) \tag{15}$$

In Equation (15), N represents the number of samples to be tested, $T(i)$ is the feature function for total accuracy.

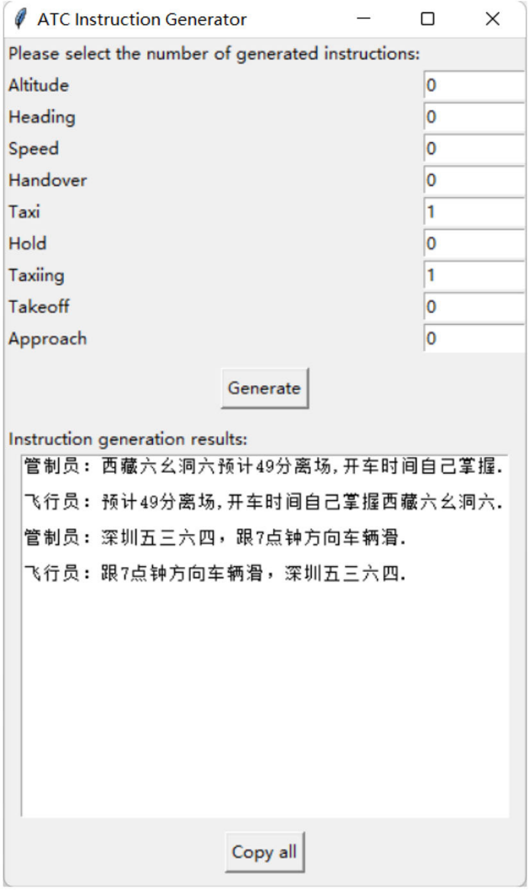
4.4. ATC Corpus Segmentation Dictionary

To facilitate the ROUGE evaluation and keyword evaluation of repetition instructions, we built a Chinese Air-Ground Communication Segmentation Dictionary based on the training data and reference the regulation "Radio Communication Phraseology for Air Traffic Services" (MH/T 4014-2003), as well as the abbreviation standards. We used the Jieba segmentation tool to construct the dictionary, which includes aviation company abbreviations, numbers, letters, altitude levels, speeds, headings, waypoints, proper nouns, and other relevant terms. The dictionary consists of a total of 14,756 vocabulary entries. A sample analysis of the vocabulary is presented in Table 2.

5. Experiments and discussions

5.1. Dataset

The experiment consists of two datasets: the area control dataset and the tower control dataset. The area control dataset comprises real air-to-ground communication data in actual ATC scenarios. The tower control dataset, on the other hand, is generated by computer based on the standards, and its User Interface (UI) is shown in Figure 4. You can find this algorithm in this link https://drive.google.com/drive/folders/1RN6CEhJXcoru6LyZB8u_Y3XBLjyvlQqd?usp=sharing. To illustrate the distribution of these two datasets, we utilized Term Frequency-Inverse Document Frequency (TF-IDF) for data vectorization and employed Principal Component Analysis (PCA) for dimensionality reduction to achieve data visualization. The dataset distributions are depicted in Figure 5.



Please select the number of generated instructions:

Altitude	0
Heading	0
Speed	0
Handover	0
Taxi	1
Hold	0
Taxiing	1
Takeoff	0
Approach	0

Instruction generation results:

管制员： 西藏六幺洞六预计49分离场,开车时间自己掌握.
飞行员： 预计49分离场,开车时间自己掌握西藏六幺洞六.
管制员：深圳五三六四，跟7点钟方向车辆滑.
飞行员：跟7点钟方向车辆滑，深圳五三六四.

FIGURE 4
UI Interface of the tower control instruction generator.

In Figure 5, the distribution represented by red stars corresponds to the area control dataset, while the distribution denoted by blue stars corresponds to the tower control dataset. It is evident that the tower control dataset encompasses a significantly different set of instruction types compared to the area control dataset, which can be used to assess the model's generalization capability.

The dataset for training the area control consists of 11,049 pairs, with 8,949 pairs used for training, 995 pairs for validation, and 1,105 pairs for testing. The tower control dataset, used for transfer learning generalization evaluation, contains a total of 1,074 pairs. Table 3 displays some examples from the dataset.

5.2. Experiment configurations

The experiments were conducted on a Windows operating system. The computer configuration is as follows: Intel Core i5-8400 processor, 56 GB of RAM, NVIDIA RTX 4090 24 GB graphics card, 250 GB SSD, and a 3.6 TB HDD. The deep

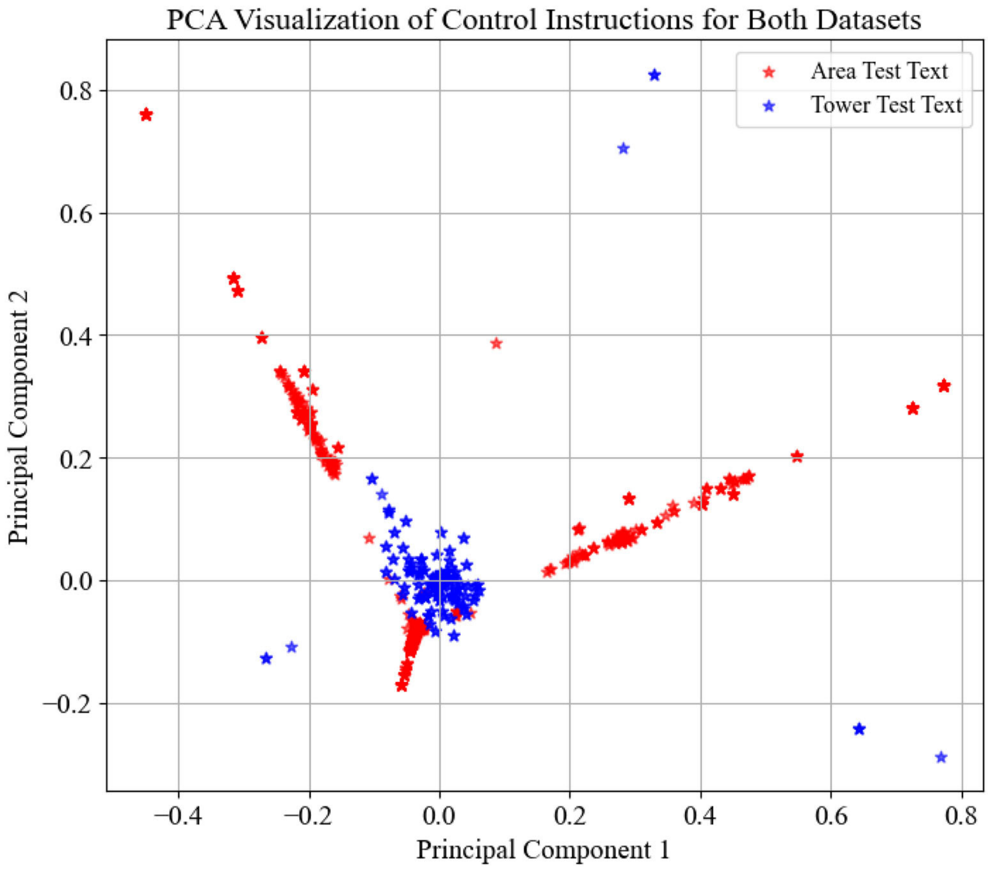


FIGURE 5
Distribution of tower control dataset and area control dataset.

TABLE 3 Dataset example table.

Dataset name	Control instructions	Pilot recitation instructions
Tower	Jinxiu 7443, estimated departure time 10 min.	Estimated departure time is 10 min, Jinxiu 7443.
	Hebei 8554, circling and waiting over JHG.	Circling and waiting over JHG, Hebei 8554.
Area	Shandong 8896, Xiamen, radar has been identified.	Radar has identified, Shandong 8896.
	Hainan 7064, cancel offset return route.	Cancel offset return route, Hainan 7064.

TABLE 4 Hyperparameters for the RoBERTa model.

Hyperparameter	Setting
Dropout	0.1
Max sequence length	256
Learning rate	0.0001
Batch size	32
Number of epochs	20
Optimizer	Adam
Beamsearch size	3
Weight decay	0.001
λ	0.5

learning framework used was PyTorch. The hyperparameters for the RoBERTa-RL model are listed in Table 4.

5.3. Ablation experiment

To demonstrate the effectiveness of the adopted strategies, we conducted ablation experiments for validation, using ROUGE-N and ROUGE-L as evaluation metrics. The experimental results are shown in Table 5.

According to Table 5, it can be observed that RoBERTa-RL($\lambda = 0$), the unimproved RoBERTa model, achieves good performance on the area control dataset through transfer learning. However, it performs poorly on the tower control dataset, indicating a problem of poor generalization when relying solely on transfer learning. When $\lambda = 0.3$, it can be seen that the model has overcome the issue of poor generalization and shows further improvement compared to $\lambda = 0$. When

TABLE 5 Experimental results based on ROUGE evaluation metrics.

Model	Dataset	ROUGE-1	ROUGE-2	ROUGE-L
RoBERTa-RL (λ = 0)	Area	0.995	0.990	0.994
	Tower	0.885	0.704	0.885
RoBERTa-RL (λ = 0.3)	Area	0.996	0.991	0.995
	Tower	0.980	0.946	0.980
RoBERTa-RL (λ = 0.5)	**Area**	**0.996**	**0.991**	**0.995**
	Tower	**0.982**	**0.954**	**0.982**
RoBERTa-RL (λ = 1.0)	Area	0	0	0
	Tower	0	0	0

The meaning of the bold values is the optimal values achieved by the RoBERTa-RL (λ = 0.5) model across different datasets and metrics.

TABLE 6 Comparative experimental results based on ROUGE evaluation metrics.

Model	Dataset	ROUGE-1	ROUGE-2	ROUGE-L
GPT2	Area	0.981	0.973	0.981
	Tower	0.779	0.61	0.776
BERT	Area	0.991	0.984	0.991
	Tower	0.846	0.662	0.846
BART	Area	0.992	0.987	0.992
	Tower	0.910	0.767	0.910
RoBERTa-RL (λ = 0)	Area	0.995	0.990	0.994
	Tower	0.885	0.704	0.885
RoBERTa-RL (λ = 0.5)	**Area**	**0.996**	**0.991**	**0.996**
	Tower	**0.982**	**0.954**	**0.982**

The meaning of the bold values is the optimal values achieved by the RoBERTa-RL (λ = 0.5) model across different datasets and metrics.

λ = 0.5, the model reaches optimal performance. This is because choosing a reward weight of 0.3 emphasizes the cross-entropy loss. On the other hand, a reward weight of 0.5 balances the contribution of the cross-entropy loss and the reward function. This setting can to some extent balance the quality and grammatical accuracy of the generated instructions, leading to better performance. Setting the reward weight λ to 1, without considering the cross-entropy loss, means only optimizing the similarity between the generated results and the reference summaries, without considering grammatical accuracy and the optimization of the generation strategy. This results in the model disregarding grammar rules and sentence structure during the generation process, leading to the generation of unreasonable instructions.

5.4. Contrastive experiments

To perform a comprehensive analysis of the constructed model's performance, we adopted a comparative research approach tailored to the application domain. Specifically, we evaluated the performance of the constructed model as well as leading pre-trained models in the field of text generation, namely GPT-2, BERT, and BART, in the task of repetition instruction generation. Tests were conducted separately on the area control dataset and the tower control dataset, with evaluation metrics including ROUGE-N, ROUGE-L, and keyword evaluation criteria. The experimental results are presented in Tables 6, 7. Furthermore, to visualize the improvements made by the model, we compiled statistics on the length distribution of repetition instructions generated by the model before and after enhancements on the tower control test dataset. The visual results are illustrated in Figures 6–8.

From Table 6, it can be observed that all comparative models performed well on the area control dataset. The proposed RoBERTa-RL(λ = 0.5) model only slightly outperformed the comparative models. However, on the tower control dataset, all comparative models showed poor generalization performance, while our proposed model's performance only slightly decreased. Table 7 provides a detailed display of the performance of each transfer learning model based on the Keyword Evaluation Metrics. From Table 7, it is visually evident that the comparative models performed poorly on the tower control dataset, indicating a clear issue of poor generalization. Additionally, the GPT-2 model performed the worst in the task, possibly due to its use of masked attention mechanism during prediction, which failed to incorporate useful information from the context. Finally, our constructed RoBERTa-RL(λ = 0.5) model achieved the best performance on the tower control dataset, demonstrating that the proposed improvement strategies greatly alleviate the issue of poor generalization in transfer learning.

TABLE 7 Comparative experimental results based on keyword evaluation metrics.

Model	Dataset	CSA (%)	AIA (%)	PA (%)	TA (%)
GPT2	Area	99.1	98.0	98.5	96.8
	Tower	89.3	81.4	24.5	23.3
BERT	Area	99.6	98.8	98.8	97.4
	Tower	**100.0**	**99.8**	25.6	25.6
BART	Area	99.2	98.8	98.2	96.8
	Tower	99.0	94.1	35.1	34.3
RoBERTa-RL (λ = 0)	Area	99.7	98.2	99.4	97.6
	Tower	98.7	94.5	25.8	25.8
RoBERTa-RL (λ = 0.5)	**Area**	**100.0**	**99.1**	**99.5**	**98.8**
	Tower	99.7	98.7	**82.5**	**81.8**

The meaning of the bold values is the optimal values achieved by the RoBERTa-RL (λ = 0.5) model across different datasets and metrics.

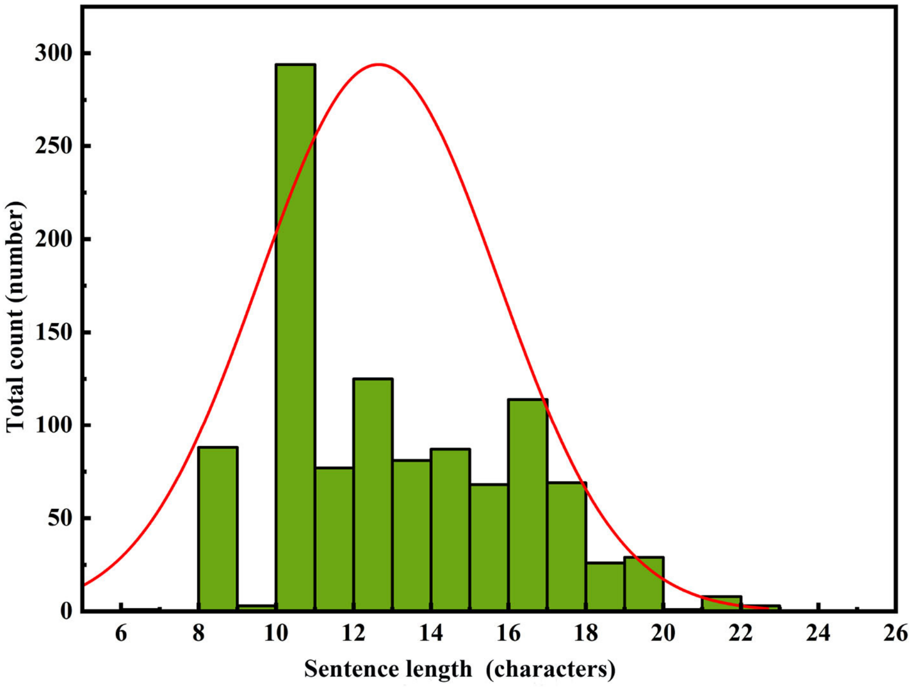


FIGURE 6
RoBERTa-RL (λ = 0) PRG text length distribution.

In Figures 6–8, the horizontal axis represents the string length of repetition instructions, while the vertical axis denotes the total count of repetition instructions of varying lengths. The red curve illustrates the length distribution of repetition instructions. By comparing Figures 6, 7, we observe that the mean length of repetition instructions generated by RoBERTa-RL (λ = 0) is lower than the mean length of reference labels, indicating a significant omission of words and poor generalization for this model. However, by comparing Figures 7, 8, we can see that the RoBERTa-RL (λ = 0.5) model generates repetition instructions with a length similar to the mean length of reference labels, effectively mitigating the omission issue and demonstrating strong generalization.

In addition, we analyzed the reasons behind the model's strong generalization capability. Specifically, due to the disparities in data distribution between the area control dataset and the tower control dataset, the baseline model fine-tuned on the area control dataset performed poorly on the tower control dataset. This generalization issue is a common challenge faced by most fine-tuned models at the current stage. However, the introduction of reinforcement learning strategies effectively mitigates this problem. During the training process, we incorporated a reward and penalty mechanism to assess the quality of generated results and provide timely feedback to the model. This mechanism encourages the model to prioritize the quality of the generated text over similarity to the target labels, thereby preventing overfitting to the training

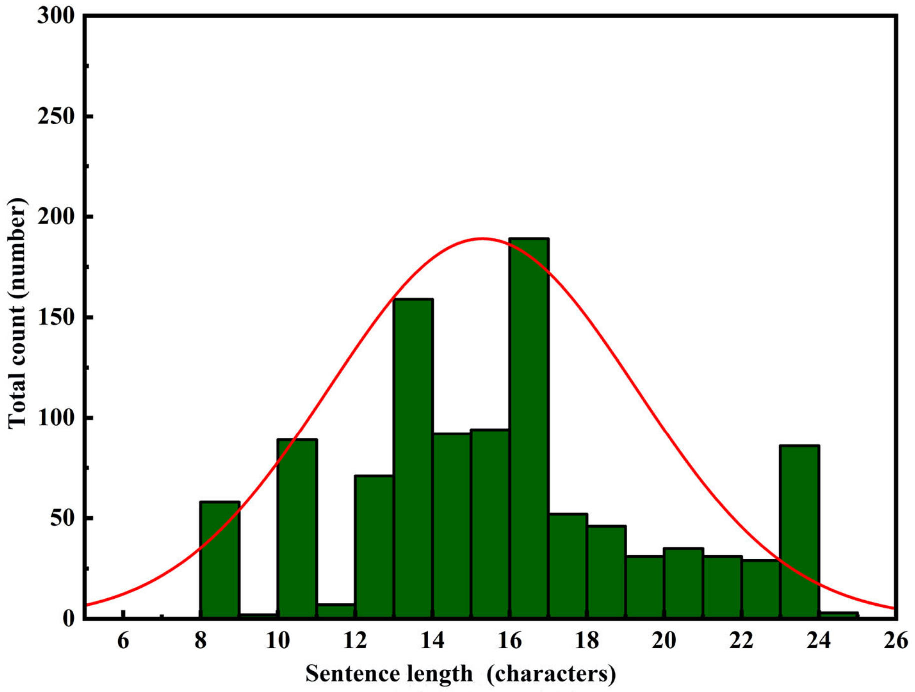


FIGURE 7
Reference label length distribution.

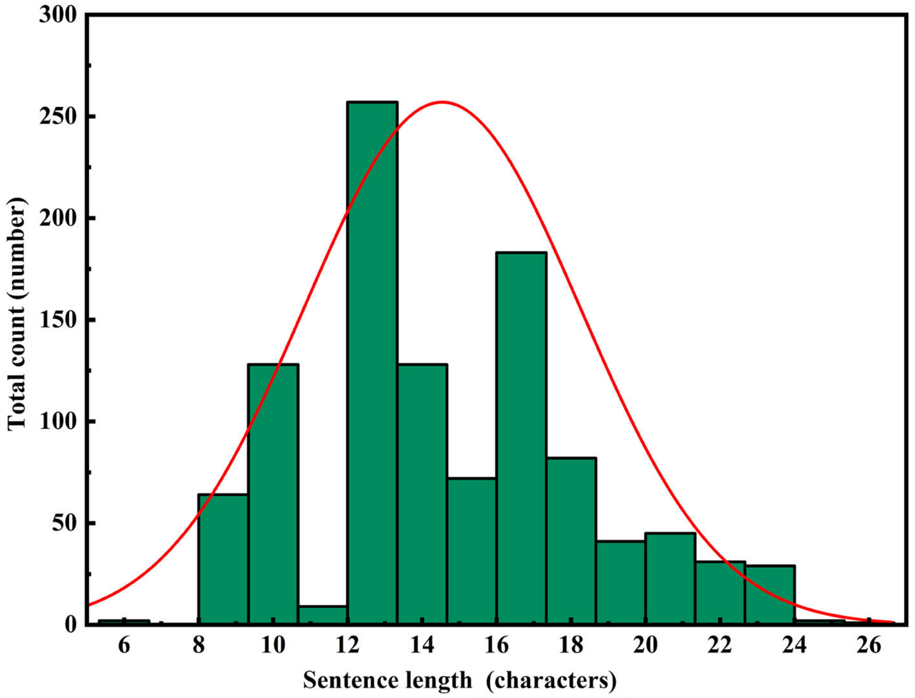


FIGURE 8
RoBERTa-RL (λ=0.5) PRG text length distribution.

data distribution. Furthermore, the introduction of the reward and penalty mechanism essentially transforms the model into a multitask learning problem, where one task is to generate repetition instructions, and the other task is to learn how to generate high-quality instructions to maximize rewards. As a result, the model's generated results exhibit strong performance on datasets with different distributions. Finally, setting the weights of both the reinforcement learning loss and the original cross-entropy loss to 0.5 ensures that the model does not overly rely on either aspect during optimization but strikes a balance between the two objectives, thereby enhancing the overall model performance. In summary, reinforcement learning strategies are advantageous in enabling the model to learn deep features of the dataset, allowing the model to excel on similar yet differently distributed datasets. This approach is highly effective and can be applied to many similar problems to improve model generalization capabilities.

6. Conclusions

Our research focuses on addressing the problem of generating high-quality pilot recitations in the ATC field based on small-scale training data. To tackle this challenge, we propose a DRL model that optimizes the cross-entropy loss using the policy gradient algorithm to overcome exposure bias and poor generalization in transfer learning. Through a series of experiments, we demonstrate that our proposed model outperforms the comparison models on the training dataset and maintains excellent performance on similar distribution datasets. To expedite model training, we employ a pretraining method based on cross-entropy loss and a training strategy that combines the policy gradient algorithm with cross-entropy loss. This strategy allows the model to converge faster and reduces resource consumption. In addition to the commonly used ROUGE evaluation metric, we introduce a keyword-based evaluation metric to assess the model's performance. The results show that the keyword-based evaluation metric provides a more accurate reflection of the model's performance. On the tower control dataset, our proposed model achieves an overall accuracy of 81.8%, which is a 56% improvement compared to the pre-improved model and a 47.5% improvement compared to the other comparable models.

However, it is essential to consider some potential safety implications that the model may introduce in practical applications. At the current stage, since the model's input is limited to textual information alone, it lacks sufficient contextual information to assess the reasonableness of the control instructions it receives. As a result, it cannot generate queries or doubts about control instructions that could lead to flight conflicts. To facilitate the deployment of the model in real-world scenarios, it is imperative that the model, in addition to processing text data, can also incorporate navigation and monitoring data. In our future work, we will integrate these multimodal data sources as inputs to the repetition generation model, enabling it to scrutinize and question conflicting or unreasonable control instructions, thereby further mitigating safety risks.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://drive.google.com/drive/folders/1RN6CEhJXcoru6LyZB8u_Y3XBLjyvlQqd?usp=sharing.

Author contributions

WP: Writing—review and editing. PJ: Writing—original draft. YL: Writing—review and editing. ZW: Writing—review and editing. JH: Writing—review and editing.

Funding

This study was co-supported by the National Key R&D Program of China (No. 2021YFF0603904), National Natural Science Foundation of China (U1733203), and Safety Capacity Building Project of Civil Aviation Administration of China (TM2019-16-1/3).

Acknowledgments

We would like to extend our heartfelt gratitude to all individuals who have contributed to this paper, excluding ourselves. Their unwavering support and invaluable assistance have significantly enriched our research work.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alexandr, N., Irina, O., Tatyana, K., Inessa, K., and Arina, P. (2021). "Fine-tuning GPT-3 for Russian text summarization," in *Data Science and Intelligent Systems: Proceedings of 5th Computational Methods in Systems and Software 2021* (Springer), 748–757.

de Souza, J. G., Kozielski, M., Mathur, P., Chang, E., Guerini, M., Negri, M., et al. (2018). "Generating e-commerce product titles and predicting their quality," in *Proceedings of the 11th International Conference on Natural Language Generation* (IOP Publishing), 233–243.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Drayton, J., and Coxhead, A. (2023). The development, evaluation and application of an aviation radiotelephony specialised technical vocabulary list. *English Specific Purposes* 69, 51–66. doi: 10.1016/j.esp.2022.10.001

Elmadani, K. N., Elgezouli, M., and Showk, A. (2020). Bert fine-tuning for Arabic text summarization. *arXiv preprint arXiv:2004.14135*.

Hastie, H. (2012). "Metrics and evaluation of spoken dialogue systems," in *Data-Driven Methods for Adaptive Spoken Dialogue Systems: Computational Learning for Conversational Interfaces* (Springer), 131–150.

Helmke, H., Ohneiser, O., Buxbaum, J., and Kern, C. (2017). "Increasing atm efficiency with assistant based speech recognition," in *Proc. of the 13th USA/Europe Air Traffic Management Research and Development Seminar* (Seattle, WA).

Helmke, H., Ohneiser, O., Mühlhausen, T., and Wies, M. (2016). "Reducing controlle workload with automatic speech recognition," in *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)* (IEEE), 1–10.

Hoekstra, J. M., and Ellerbroek, J. (2016). "Bluesky atc simulator project: an open data and open source approach," in *Proceedings of the 7th International Conference on Research in Air Transportation*, 132.

Holone, H., and Nguyen, V. N. (2015). Possibilities, challenges and the state of the art of automatic speech recognition in air traffic control. *Int. J. Comput. Inform. Eng.* 9, 1933–1942.

Keneshloo, Y., Ramakrishnan, N., and Reddy, C. K. (2019). "Deep transfer reinforcement learning for text summarization," in *Proceedings of the 2019 SIAM International Conference on Data Mining* (SIAM), 675–683.

Kleinert, M., Helmke, H., Siol, G., Ehr, H., Finke, M., Srinivasamurthy, A., et al. (2017). "Machine learning of controller command prediction models from recorded radar data and controller speech utterances," in *7th SESAR Innovation Days* (Belgrade).

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., et al. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Lin, C.-Y., and Och, F. (2004). "Looking for a few good metrics: rouge and its evaluation," in *NTCIR Workshop*.

Lin, Y. (2015). Spoken instruction understanding in air traffic control: challenge, technique, and application. *Aerospace* 8, 65. doi: 10.3390/aerospace8030065

Lin, Y., Deng, L., Chen, Z., Wu, X., Zhang, J., and Yang, B. (2019). A real-time ATC safety monitoring framework using a deep learning approach. *IEEE Trans. Intell. Transport. Syst.* 21, 4572–4581. doi: 10.1109/TITS.2019.2940992

Lin, Y., Wu, Y., Guo, D., Zhang, P., Yin, C., Yang, B., et al. (2021). A deep learning framework of autonomous pilot agent for air traffic controller training. *IEEE Trans. Hum. Mach. Syst.* 51, 442–450. doi: 10.1109/THMS.2021.3102827

Liu, M., Wang, Z., and Wang, L. (2021). Automatic Chinese text summarization for emergency domain. *J. Phys. Conf. Ser.* 1754, 012213. doi: 10.1088/1742-6596/1754/1/012213

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). RoBERTa: a robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Luo, Z., Huang, S., and Zhu, K. Q. (2019). Knowledge empowered prominent aspect extraction from product reviews. *Inform. Process. Manage.* 56, 408–423. doi: 10.1016/j.ipm.2018.11.006

Mane, M. R., Kedia, S., Mantha, A., Guo, S., and Achan, K. (2020). Product title generation for conversational systems using BERT. *arXiv preprint arXiv:2007.11768*.

Miao, L., Cao, D., Li, J., and Guan, W. (2020). Multi-modal product title compression. *Inform. Process. Manage.* 57, 102123. doi: 10.1016/j.ipm.2019.102123

Nallapati, R., Xiang, B., and Zhou, B. (2016a). "Sequence-to-sequence rnns for text summarization," in *Workshop Track - ICLR 2016*.

Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B., et al. (2016b). Abstractive text summarization using sequence-to-sequence RNNs and beyond. *arXiv preprint arXiv:1602.06023*.

Nazari, N., and Mahdavi, M. (2019). A survey on automatic text summarization. *J. AI Data Mining* 7, 121–135.

Pan, W., Jiang, P., Wang, Z., Li, Y., and Liao, Z. (2023). Ernie-gram biGRU attention: an improved multi-intention recognition model for air traffic control. *Aerospace* 10, 349. doi: 10.3390/aerospace10040349

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI Blog* 1, 9.

Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

See, A., Liu, P. J., and Manning, C. (2017). Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Song, S., Huang, H., and Ruan, T. (2019). Abstractive text summarization using LSTM-CNN based deep learning. *Multimedia Tools Appl.* 78, 857–875. doi: 10.1007/s11042-018-5749-3

Srinivasamurthy, A., Motlicek, P., Himawan, I., Szaszak, G., Oualil, Y., and Helmke, H. (2017). "Semi-supervised learning with semantic knowledge extraction for improved speech recognition in air traffic control," in *Proceedings of Interspeech 2017*, 2406–2410.

Sun, G., Wang, Z., and Zhao, J. (2021). Automatic text summarization using deep reinforcement learning and beyond. *Inform. Technol. Control* 50, 458–469. doi: 10.5755/j01.itc.50.3.28047

Yang, B., Tan, X., Chen, Z., Wang, B., Li, D., Yang, Z., et al. (2019). ATCspeech: a multilingual pilot-controller speech corpus from real air traffic control environment. *arXiv preprint arXiv:1911.11365*.

Yiu, C. Y., Ng, K. K., Lee, C.-H., Chow, C. T., Chan, T. C., Li, K. C., et al. (2021). A digital twin-based platform towards intelligent automation with virtual counterparts of flight and air traffic control operations. *Appl. Sci.* 11, 10923. doi: 10.3390/app112210923

Zhang, J., Zhang, P., Guo, D., Zhou, Y., Wu, Y., Yang, B., et al. (2022a). Automatic repetition instruction generation for air traffic control training using multi-task learning with an improved copy network. *Knowledge Based Syst.* 241, 108232. doi: 10.1016/j.knosys.2022.108232

Zhang, M., Zhou, G., Yu, W., Huang, N., and Liu, W. (2022b). A comprehensive survey of abstractive text summarization based on deep learning. *Comput. Intell. Neurosci.* 2022, 7132226. doi: 10.1155/2022/7132226

Zhang, Y., Merck, D., Tsai, E. B., Manning, C. D., and Langlotz, C. P. (2019). Optimizing the factual correctness of a summary: a study of summarizing radiology reports. *arXiv preprint arXiv:1911.02541*.

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms*. CRC Press.

Zuluaga-Gomez, J., Prasad, A., Nigmatulina, I., Motlicek, P., and Kleinert, M. (2023). A virtual simulation-pilot agent for training of air traffic controllers. *Aerospace* 10, 490. doi: 10.3390/aerospace10050490

frontiers | Frontiers in Neurorobotics

TYPE Original Research
PUBLISHED 12 October 2023
DOI 10.3389/fnbot.2023.1276519

Check for updates

OPEN ACCESS

EDITED BY
Long Jin,
Lanzhou University, China

REVIEWED BY
Yuxuan Cao,
Naval University of Engineering, China
Yinyan Zhang,
Jinan University, China

*CORRESPONDENCE
Wanruo Zhang
✉ wanruo_zhang@stu.pku.edu.cn

RECEIVED 12 August 2023
ACCEPTED 19 September 2023
PUBLISHED 12 October 2023

CITATION
Xu J, Zhang W, Cai J and Liu H (2023) SafeCrowdNav: safety evaluation of robot crowd navigation in complex scenes. *Front. Neurorobot.* 17:1276519. doi: 10.3389/fnbot.2023.1276519

SafeCrowdNav: safety evaluation of robot crowd navigation in complex scenes

Jing Xu[1,2], Wanruo Zhang[1]*, Jialun Cai[1] and Hong Liu[1]

[1]Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, Shenzhen, China,
[2]School of Computer Science and Technology, Xidian University, Xi'an, China

Navigating safely and efficiently in dense crowds remains a challenging problem for mobile robots. The interaction mechanisms involved in collision avoidance require robots to exhibit active and foresighted behaviors while understanding the crowd dynamics. Deep reinforcement learning methods have shown superior performance compared to model-based approaches. However, existing methods lack an intuitive and quantitative safety evaluation for agents, and they may potentially trap agents in local optima during training, hindering their ability to learn optimal strategies. In addition, sparse reward problems further compound these limitations. To address these challenges, we propose SafeCrowdNav, a comprehensive crowd navigation algorithm that emphasizes obstacle avoidance in complex environments. Our approach incorporates a safety evaluation function to quantitatively assess the current safety score and an intrinsic exploration reward to balance exploration and exploitation based on scene constraints. By combining prioritized experience replay and hindsight experience replay techniques, our model effectively learns the optimal navigation policy in crowded environments. Experimental outcomes reveal that our approach enables robots to improve crowd comprehension during navigation, resulting in reduced collision probabilities and shorter navigation times compared to state-of-the-art algorithms. Our code is available at https://github.com/Janet-xujing-1216/SafeCrowdNav.

KEYWORDS

mobile robot, human-aware navigation, reinforcement learning, security assessment, collision avoidance

1. Introduction

Mobile robots have been extensively studied and widely applied in recent decades as an essential branch of robotics research. They can accomplish tasks that are difficult or impossible for humans, reduce the workload of human workers, and improve people's quality of life. Our daily lives increasingly depend on mobile robots, which share living and social spaces with humans and interact with them to varying degrees. The crucial factor determining the successful autonomous movement of mobile robots across diverse environments is their possession of adaptable and autonomous navigation capabilities.

The key to achieving efficient autonomous navigation of mobile robots in various environments lies in key elements such as safety, autonomy, effectiveness, and user-friendliness. Among these, obstacle avoidance (Duguleana and Mogan, 2016; Pandey et al., 2017), serving as a primary means to ensure safety, poses a challenging research problem in robot navigation. It has been studied for decades and finds applications in critical real-world scenarios such as autonomous driving (Kästner et al., 2021) and cargo logistics. For instance, in the context of mobile robots, scenarios like autonomous navigation within unmanned supermarkets or warehouses, where robots navigate among shoppers or workers while avoiding obstacles, have garnered significant attention. At the same time, the operating environments for mobile robots have become increasingly complex, with various static and dynamic obstacles coexisting, including obstacles such as barriers, pedestrians, vehicles, or other robots. These scenarios add a layer of complexity, as robots must safely maneuver in dynamic environments alongside pedestrians and other obstacles, showcasing the versatility and practicality of mobile robotics. While classical planning methods (Cai et al., 2023) can effectively handle static environments, reliable obstacle avoidance in dynamic environments remains a significant challenge. Safe and reliable navigation in these highly dynamic environments is still a crucial challenge.

The illustration of our work is showing in Figure 1 and the paper presents the following key contributions:

- We design a novel framework called SafeCrowdNav, which integrates hindsight experience replay and prioritized experience replay to address the challenge of sparse-reward navigation.
- We firstly propose novel safety evaluation reward functions to estimate the safety weights of the robot in its current state, enabling more accurate obstacle avoidance during the navigation process.
- We firstly propose a novel intrinsic exploration reward function with visited count state that helps the robot avoid getting stuck in place and reduces unnatural robot behavior.

2. Related works

2.1. React-based collision avoidance

Over the past decade, extensive research has focused on robotic navigation in dynamic obstacle environments within the field of robotics. Numerous works have been dedicated to classical navigation techniques, with the earliest attempts being reactive rules-based methods, such as Optimal Reciprocal Collision Avoidance (ORCA) (Van den Berg et al., 2008), Reciprocal Velocity Obstacle (RVO) (Van Den Berg et al., 2011), and Social Force (SF) (Helbing and Molnar, 1995). These methods employ one-step interaction rules to determine the robot's optimal actions. However, despite considering interactions among agents, ORCA and SF simplify the crowd behavior model, leading to limitations such as shortsightedness, lack of safety, and unnatural movement patterns.

2.2. Trajectory-based collision avoidance

As a result, researchers have started exploring trajectory-based methods (Kothari et al., 2021) and considered visual-inertial initialization (Huang et al., 2021; Liu et al., 2022) to address crowd avoidance problems. Nevertheless, trajectory-based approaches suffer from high computational costs, inability to perform real-time updates in the presence of increasing crowd sizes and difficulties in finding safe paths (Trautman and Krause, 2010; Alahi et al., 2016; Sathyamoorthy et al., 2020). These limitations restrict the application and effectiveness of these methods in large-scale crowd scenarios.

2.3. Learning-based collision avoidance

To overcome the above challenges, recent research has modeled the crowd navigation problem as a Markov Decision Process (MDP) and introduced deep reinforcement learning called Collision Avoidance with Deep Reinforcement Learning (CADRL). Chen et al. (2019) propose the Socially Attentive Reinforcement Learning (SARL), which combines human-robot interaction features with self-attention mechanisms to infer the relative importance of neighboring humans with respect to their future states. They also develop the simulation environment CrowdNav (Chen et al., 2019), which has been widely used for comparing CADRL approaches. In CrowdNav, the information regarding the agent's position, velocity, and radius is considered as input, and the robot responds accordingly based on this input. To address the computational cost associated with learning-based methods, Zhou et al. (2022) propose SG-D3QN, which utilizes graph convolutional networks to predict social attention weights and refines coarse Q-values through online planning of potential future trajectories. The latest paper (Martinez-Baselga et al., 2023) claims to be the first work in this field that applies intrinsic rewards and has achieved the state-of-the-art performance.

2.4. Safety evaluation

However, reinforcement learning algorithms suffer from a fatal drawback: the need for trial and error exploration of the environment to learn optimal policies. In real-world settings, safety is a crucial concern, and trial and error that may cause harm to humans during the exploration process is unacceptable. Although current practices often train reinforcement learning agents in simulation environments with low safety risks, the complexity of transitioning from simulated environments to the real world poses a series of unacceptable safety issues (Ray et al., 2019). Therefore, safety evaluation should be a key focus area in reinforcement learning research. In this regard, this paper is dedicated to addressing safety concerns and proposes a robot crowd navigation system that enables the evaluation of an agent's safety performance.

3. Problem formulation

3.1. Crowd navigation modeling

The problem of crowd navigation for robots refers to guiding a robot to its target location in the shortest possible time while avoiding collisions with a variable number of intelligent agents behaving like a crowd in the environment. These agents can encompass various types of obstacles, and in this study, we utilize the CrowdNav simulation environment widely adopted in previous works (Chen et al., 2019, 2020; Everett et al., 2021).

The observable state of all agents w is represented by their positions $p = [p_x, p_y]$, velocities $v = [v_x, v_y]$, and radii r. The

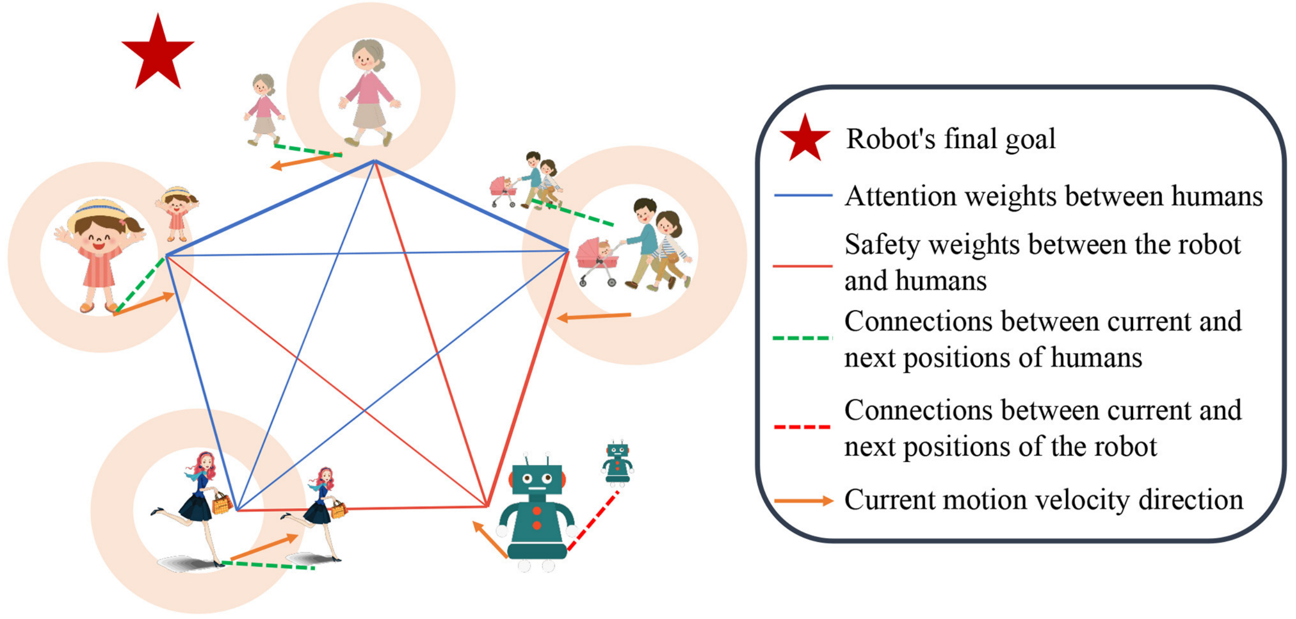

Robot's final goal
Attention weights between humans
Safety weights between the robot and humans
Connections between current and next positions of humans
Connections between current and next positions of the robot
Current motion velocity direction

FIGURE 1
Illustration of our work: the robot utilizes heterogeneous attention weights and safety evaluation scores obtained from observations to selectively aggregate pedestrian information, enabling more anticipatory decision-making.

observable state indicates the information that other visible agents in the environment can perceive. Additionally, the state of the robot includes its preferred velocity (v_p), heading angle (θ), and target coordinates ($g = [g_x, g_y]$). At a given time step t, the input joint state of the robot s^t is defined as:

$$\begin{aligned} s^t &= \left[w_r^t, w_h\right] \\ w_r^t &= \left[p_x^t, p_y^t, v_x^t, v_y^t, r^t, g_x^t, g_y^t, v_p^t, \theta^t\right] \\ w_h &= \left[w_1^t, w_2^t, \ldots, w_n^t\right] \\ w_i^t &= \left[p_x^i, p_y^i, v_x^i, v_y^i, r\right], i > 0, \end{aligned} \tag{1}$$

where w_r^t is the state of the robot r, w_i^t is the state of human agent i and w_h is the collective state of all human agents.

3.2. Reinforcement learning based on the Q-value

In our work, the crowd navigation problem is formulated as a Markov Decision Process, and we adopt the double dueling deep Q-network as the fundamental method for solving this task. The objective is to estimate the optimal policy π^*, which selects the optimal action a^t for state s^t at a specific time step t. The optimal policy maximizes the expected return, given by:

$$\pi^*\left(s^t\right) = \underset{a^t}{\operatorname{argmax}}\left(Q^*\left(s^t, a^t\right)\right), \tag{2}$$

where Q^* is the optimal action-value function, recursively defined with the Bellman equation as:

$$Q^*\left(s^t, a^t\right) = \mathbb{E}\left[r^t + \gamma^{\Delta t \cdot v_p} \max_{a^{t+1}} Q^*\left(s^{t+1}, a^{t+1}\right)\right], \tag{3}$$

where s^{t+1} is the successor state and r^t is immediate reward. $\gamma \in (0, 1)$ is the discount factor that balances the current and future rewards, normalized by the preferred velocity v_p and the time step size Δt.

3.3. Reward shaping

While tackling the challenge of sparse reward tasks in crowd navigation without expert demonstrations, the most intuitive approach is to shape the reward function. However, previous works (Chen et al., 2017, 2019) have not given due attention to this aspect and instead applied sparse reward functions designed for non-communicative dyadic collision avoidance problems. In crowd navigation, such mismatched rewards can lead to poor training convergence (Chen et al., 2020). In contrast to existing reward functions (Chen et al., 2019; Zhou et al., 2022), which commonly rely solely on external or intrinsic rewards, our approach not only integrates and refines these two reward functions, but also introduces an additional safety evaluation function. We divide the overall reward r^t into three parts and innovate each: externally provided rewards r_{ex}^t, safety evaluation function r_{safe}^t, and intrinsic exploration rewards r_{in}^t, defined as follows:

$$r^t = r_{ex}^t + r_{safe}^t + r_{in}^t, \tag{4}$$

where we first introduce innovations in the externally-provided reward function r_{ex}^t offered by the environment to incentivize the robot to navigate toward the goal while avoiding collisions. Additionally, we introduced safety evaluation functions r_{safe}^t and intrinsic rewards r_{in}^t to encourage the robot to explore and exploit the environment while improving its safety and reliability.

4. Method

This paper focuses on the safety evaluation of crowd navigation using deep reinforcement learning. Building upon SG-D3QN (Zhou et al., 2022), we firstly model the social relationship graph (Liu et al., 2023), a heterogeneous spatio-temporal graph as input to the SG-D3QN planner to generate optimal actions. The simulated environment provides external reward function, safety evaluation scores and intrinsic exploration reward function based on the current state, which are then fed back to the reinforcement learning policy. The trajectory sampling process combines hindsight experience replay and prioritized experience replay to handle the data in the experience replay buffer. The overall framework of our algorithm is illustrated in Figure 2.

4.1. External reward function

We redesign the external reward function r_{ex}^t offered by the environment, dividing it into r_{goal}^t, $r_{collision}^t$, $r_{shaping}^t$, r_{pred}^t four components. r_{goal}^t is used to reward the robot for reaching the goal, $r_{collision}^t$ penalizes collisions, $r_{shaping}^t$ guides the robot toward the goal, and r_{pred}^t provides penalties for potential collisions in future time steps. Our external reward function is defined as follows:

$$r_{ex}^t = r_{goal}^t + r_{collision}^t + r_{shaping}^t + r_{pred}^t. \tag{5}$$

The individual components r_{goal}^t, $r_{collision}^t$, $r_{shaping}^t$, r_{pred}^t are defined as follows:

$$r_{\mathrm{goal}}^t = \begin{cases} r_{\mathrm{arr}} & \text{if target is reached} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

$$r_{\mathrm{collision}}^t = \begin{cases} r_{\mathrm{col}} & \text{if collision} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

$$r_{\mathrm{shaping}}^t = w_p \cdot \left(\| p^{t-1} - p_g \| - \| p^t - p_g \| \right) \tag{8}$$

$$r_{\mathrm{pred}}^t = \min_{i=1,\ldots,n} r_{\mathrm{pred}}^{i,t} = \min_{i=1,\ldots,n} \left[\min_{k=1,\ldots,K} \left(\mathbb{1}_i^{t+k} \frac{r_{\mathrm{col}}}{2^k} \right) \right], \tag{9}$$

where r_{shaping}^t represents the difference between the distance from the endpoint at time $t-1$ and t. p^t and p_g respectively represent the robot's position and the goal at time t, and w_p is a hyper-parameter. Prediction reward function r_{pred}^t presents the maximum penalty for collisions occurring among n humans in future K time steps. $\mathbb{1}_i^{t+k}$ indicates whether the robot collides with the predicted position of the human i at time $t+k$. The role of 2^k is to assign different weights to collisions at different predicted time steps, with lower penalty weights given to collisions predicted farther into the future.

4.2. Safety evaluation function

The safety evaluation function r_{safe}^t assesses the current safety level of the robot based on the surrounding environment information and adjusts the robot's behavior accordingly to guide it toward safer navigation. Specifically, if the safety evaluation function r_{safe}^t provides a higher safety score, it indicates a lower risk and likelihood of collisions in the current environment, allowing the robot to choose a relatively higher speed to complete the navigation task more quickly. Conversely, if the safety evaluation function r_{safe}^t provides a lower safety score, it indicates a higher risk and likelihood of collisions in the current environment, requiring the robot to lower its speed or even stop to avoid potential danger. The factors considered in the safety evaluation function include:
(1) Collision probability $r_{obstacle}^t$ between the robot and obstacles: It considers the movement speed and direction of obstacles, the distance between the robot and obstacles, and the obstacle type together. A global collision probability map is used here, where closer obstacles to the robot have a higher collision probability $p_{\mathrm{collision}}$.
(2) Robot's velocity r_{robot}^t: Ensuring smooth and natural motion is vital in dynamic and crowded settings, enhancing comfort and safety for passengers and bystanders. Abrupt velocity changes can cause discomfort and confusion among humans and destabilize navigation, leading to collisions. Thus, we quantify motion smoothness by assessing continuity in velocity changes, calculated from the cosine of the angle between current v^t and previous v^{t-1} robot actions.
(3) Safety distance $r_{discomfort}^t$ between obstacles and the robot: To ensure the safety and comfort of humans during robot navigation, we additionally impose a penalty when the distance between obstacles and the robot falls below the predefined safety threshold. Actually, collision probability $r_{obstacle}^t$ can partially achieve this goal, but only use it fail to discourage situations that may potentially cause discomfort to humans.

The composition of the safety score is as follows:

$$r_{safe}^t = r_{obstacle}^t + r_{robot}^t + r_{discomfort}^t \tag{10}$$

$$r_{\mathrm{obstacle}}^t = \beta \cdot p_{\mathrm{collision}} \tag{11}$$

$$r_{robot}^t = \alpha \cdot \frac{\overrightarrow{v^{t-1}} \cdot \overrightarrow{v^t}}{\left|\overrightarrow{v^{t-1}}\right| \, |\overrightarrow{v^t}|} \tag{12}$$

$$\begin{aligned} r_{discomfort}^t &= \sum_{i=1}^{N} f\left(d_i^t, d_s\right) \\ f\left(d_i^t, d_s\right) &= \begin{cases} d_i^t - d_s & \text{if } d_i^t < 0.2 \\ 0 & \text{else} \end{cases}, \end{aligned} \tag{13}$$

where β is a hyper-parameter, $p_{\mathrm{collision}}$ is our collision probability and v^t represents the velocity of the robot at the current time step t. Discomfort reward function $r_{discomfort}^t$ encourages the robot to maintain a safe distance from all pedestrians, where d_s is the minimum safe distance that the robot needs to maintain with pedestrians at any time. In this paper, d_s is set to 0.2 m, d_i^t represents the actual minimum distance between the robot and the i-th pedestrian within the time step.

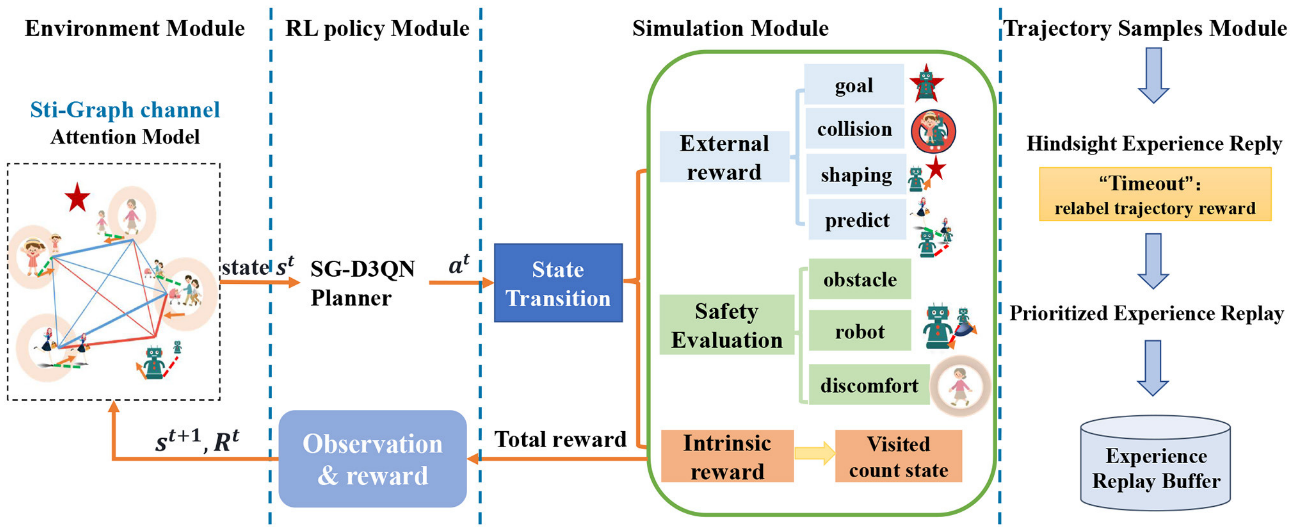


FIGURE 2
Architecture of SafeCrowdNav: **Environment Module**: Models the current environment information as a heterogeneous spatiotemporal graph. **RL Policy Module**: Implements an online planner based on SG-D3QN. Takes the state information as input and outputs the optimal action. **Simulation Module**: Divides the reward function in the reinforcement learning policy into three parts: intrinsic exploration reward, safety evaluation, and extrinsic reward. Optimizes the reward function. **Trajectory Sampling Module**: Combines hindsight experience replay and prioritized experience replay. Adjusts the reward for failed trajectories and performs experience importance sampling.

Inspired by Wang et al. (2022), our collision probability $p_{\text{collision}}$ is:

$$p_{\text{collision}} = \sum_{(x,y)\in\phi_{\text{human}}}^{i=1,\ldots,n} g_i(x,y), \tag{14}$$

where ϕ_{human} represents the range of human perception, determined by the velocities of the robot and humans and the unit of time. $g_i(x,y)$ denotes the collision probability of the robot relative to human i. "Arrive" refers to the distance between the agent and its target position being less than 0.1 m. At time t, $g_i(x,y)$ can be computed as follows:

$$g_i^t\left(x^t,y^t\right) = \sum_{i=1}^{N} N(\delta_x,x)\cdot N(\delta_y,y)\cdot N(\delta_\theta,\theta) \tag{15}$$

$$N(\delta,a) = \frac{\delta}{\sqrt{2\pi}}e^{-\frac{\left(a^t-a_i^o\right)^2}{2}} \tag{16}$$

$$\theta_i^o = \arctan\left(\frac{v_i^y}{v_i^x}\right) \qquad \theta^t = \arctan\left(\frac{y^t-y_i^o}{x^t-x_i^o}\right), \tag{17}$$

where N is the number of obstacles, and δ_x, δ_y, and δ_z are hyper-parameters representing variances. (x_i^o, y_i^o) represents the position of obstacle i, and θ_i^o denotes the heading angle of obstacle i. θ^t is the angle between the line from the robot's position (x^t, y^t) to the obstacle *is* position (x_i^o, y_i^o) and the x-axis.

Finally, the safety scores are introduced to assess the safety of the current environment. Based on these scores, the robot's behavior is modified to navigate and avoid collisions with the crowd. This approach aims to reduce the risk of collision by providing real-time analysis and guidance in response to the assessed safety levels.

4.3. Intrinsic reward function

The intrinsic reward encourages the robot to explore new states or reduce the uncertainty of predicted action outcomes (Badia et al., 2020). In this work, the intrinsic reward incentivizes the agent to visit unknown or unpredictable states until they are adequately explored and exploited, particularly in the vicinity of humans and the goal. Incorporating intrinsic exploration is beneficial in this context. Our approach is based on the Intrinsic Curiosity Module (ICM) (Pathak et al., 2017).

First, the states s and next states s_{t+1} are encoded as inputs to the feature encoder network ϕ, resulting in feature representations in the feature space $\phi(s_t)$ and $\phi(s_{t+1})$. This step aims to transform the agent-level states into state representations defined by feature vectors as outputs of the feature encoder network. Then, the states in the feature space are used to predict the actions taken, denoted as $\hat{a}_t$. Simultaneously, the actual actions a and the feature space states $\phi(s_t)$ are used to predict the next states in the feature space $\hat{\phi}\left(s_{t+1}\right)$. We adopt the same feature encoder network as (Martinez-Baselga et al., 2023), and the intrinsic reward is calculated as the mean squared error (MSE) between $\phi\left(s_{t+1}\right)$ and $\hat{\phi}\left(s_{t+1}\right)$, where higher MSE indicates that the agent is accessing unknown or unpredictable states.

To tackle the challenge of inefficient navigation resulting from excessive exploration, such as repetitive behavior within the same area, we have incorporated a state visitation record mechanism. This enhancement optimizes the exploration strategy and effectively curbs trajectory loops. The intrinsic reward r_{in} is formulated as follows:

$$r_{\text{in}} = \mu\frac{\text{MSE}\left(\phi\left(s_{t+1}\right),\hat{\phi}\left(s_{t+1}\right)\right)}{\sqrt{C\left(s_{t+1}\right)}}, \tag{18}$$

where μ is a hyper-parameter and $C(s_t)$ represents the visited count of states at time step t, indicating the number of times the robot has observed state s_t. The visited count is used to drive the robot out of already visited areas to avoid trajectory loops in the same region. The visited count state is computed on a per-episode basis, $C_{ep}(s_t) = C(s_t)$.

4.4. Experience replay

Traditional experience replay algorithms only store the experiences generated by the interaction between the agent and the environment (i.e., state, action, reward, and next state) and randomly sample them for training the agent. However, these approaches overlook valuable information, such as the agent's erroneous decisions and the significance of experiences. Errors in decision-making provide valuable learning opportunities for agents to improve their future actions, while the significance of experiences helps prioritize the replay of important events, allowing agents to learn more efficiently from crucial interactions. Therefore, we propose combining the prioritized experience replay and hindsight experience replay algorithms.

The key advantage of Prioritized Experience Replay (PER) (Schaul et al., 2015) lies in its ability to prioritize and sample important experiences, thereby enabling more effective utilization of the agent's training data. PER introduces a priority queue that efficiently sorts experiences based on their importance for training the agent, giving higher priority to experiences that are more beneficial for training. The sampling probability, denoted as $P(i)$, is monotonic with respect to the priority of the transition, ensuring a non-zero probability even for transitions with the lowest priority. In our approach, we adopt the rank-based prioritization sampling method $p(i)$ in order to enhance robustness and reduce sensitivity to outliers:

$$P(i) = \frac{p_i^{\alpha}}{\sum_k p_k^{\alpha}} \tag{19}$$

$$p_i = \frac{1}{\mathrm{rank}(i)}, \tag{20}$$

where α is a hyper-parameter that determines the degree of prioritization in the sampling and controls the exponentiation of the priorities p_i in the calculation of the sampling probabilities $P(i)$. Higher values of α emphasize experiences with higher priorities, enabling a more focused exploration of important experiences during replay.

Hindsight Experience Replay (HER) (Andrychowicz et al., 2017) addresses the specific case of failed experiences. While traditional experience replay algorithms overlook valuable information gained from failed experiences, HER can transform failed experiences into successful ones and add them to the experience replay buffer, thus effectively leveraging the knowledge from unsuccessful attempts. The key idea is to treat the final state as an additional goal, allowing the agent to learn useful information from failed simulated trajectories as if the agent had intended to reach that state from the beginning.

We present enhancements to the proposed algorithm (Li et al., 2021) tailored to suit our specific task better. Specifically, when a collision occurs or the agent reaches the goal in each episode, we store the trajectory in the experience replay buffer. If the agent's final state exceeds the global time limit ("Timeout") without causing discomfort to humans (i.e., the shortest distance is less than the safety distance), we relabel the final state as reaching the goal and assign the last reward as half of the success reward. The modified trajectory is then stored in the replay buffer. The HER method is a straightforward approach without complex reward engineering, contributing to improved sample efficiency in reinforcement learning. The details of the HER algorithm are outlined in Algorithm 1.

```
Output: experience replay memory $E$
Initialize value network $V$ and target value
 network $\hat{V}$
Initialize experience replay memory $E$
for $episode = 1$ to $M$ do
    Sample an initial state $s_0$ with the original
     goal $g$
    for $t = 1$ to $T - 1$ do
        $a^t \leftarrow \pi^*(s^t) = \underset{a^t}{\mathrm{argmax}}\left(Q^*(s^t, a^t)\right) =$
         $\mathbb{E}\left[r^t + \gamma^{\Delta t \cdot v_p} \max_{a^{t+1}} Q^*(s^{t+1}, a^{t+1})\right]$
        Execute the action $a^t$ and observe a new
         state $s^{t+1}$
    Record information info of the last state $s^T$
    if $info = ReachGoal$ or $Collision$ then
        for $t = 1$ to $T - 1$ do
            Store the transition $(s^t, a^t, r^t, s^{t+1})$ in $E$
    else if $info = Timeout$ then
        Relabel the final agent position as the
         additional goal: $g' \leftarrow p^T$
        for $t = 1$ to $T - 1$ do
            Obtain the goals $s^t_{new}$ and $s^{t+1}_{new}$ with the new
             goal $g'$;
            if $p^t = g'$ then $r^t_{new} = 1$;
            else $r^t_{new} = r^t$;
            Store the transition $(s^t_{new}, a^t, r^t_{new}, s^{t+1}_{new})$ in $E$
    for $t = 0$ to $N$ do
        Sample a minibatch $B$ from $E$ with prioritized
         sampling
        Calculate importance sampling weights
         $w_i = \left(\frac{1}{N \cdot p_i}\right)^{\beta}$
        Normalize the importance sampling weights
         $w_i = \frac{w_i}{\max(w)}$
        Compute TD errors $\delta$
        Update priorities in $E$ based on the TD
         errors
        Set target $y_i = r^t + \gamma^{\Delta t \cdot v_p} \max_{a^{t+1}} Q^*(s^{t+1}, a^{t+1})$
        Update value network $V$ by gradient descent
         with the weighted loss
    if $episode \% target update interval = 0$ then
        Update target network $V \leftarrow V'$
```

Algorithm 1. D3QN with HER and PER algorithm.

TABLE 1 Quantitative results: "Success:" the rate of the robot reaching its goal without a collision. "Collision:" the rate of the robot colliding with other humans. "Nav. Time:" the robot's navigation time to reach its goal in seconds. "Avg. Return:" discounted cumulative reward in a navigation task.

Method	*Successs*↑	*Collision*↓	*Nav. Time*↓	*Avg. Return*↑
OCRA (Van den Berg et al., 2008)	0.736	0.252	13.865	0.3234
AEMCARL (Wang et al., 2022)	0.920	0.045	12.859	0.5392
Intrinsic-SGD3QN (Martinez-Baselga et al., 2023)	0.966	0.034	**9.793**	0.6964
Hindsight & prioritized experience reply (ours)	0.948	0.052	11.753	0.6194
Intrinsic-Ntimes (ours)	0.977	0.023	10.036	0.7028
Experience reply & intrinsic-Ntimes (ours)	0.980	0.019	10.282	0.6953
SafeCrowdNav(ours)	**0.986**	**0.014**	9.984	**0.7070**

Bold values indicate the best performance of four metric.

5. Experiments

5.1. Implementation details

This paper uses Open-Gym to create a simulation environment for modeling crowd behavior and conducting path planning. Specifically, we build upon the commonly used CrowdNav simulation environment (Chen et al., 2019), which simulates crowd behavior in indoor scenarios. It incorporates factors such as crowd density and movement directions, enabling us to better study crowd behavior and path planning problems, as well as facilitating algorithm comparison.

Within each scene of the CrowdNav environment, we set up five dynamic obstacles within a circular area, requiring them to pass through the center of the circle. In more complex scenarios, we add five randomly placed individuals who must traverse the room. They navigate using the ORCA (Van den Berg et al., 2008) algorithm to avoid collisions with each other. The robot is invisible to them, meaning pedestrians in the simulation will never yield to it. This necessitates the robot to have a more proactive and anticipatory collision avoidance strategy, requiring it to execute complete obstacle avoidance maneuvers. When one person reaches a specified goal, another goal is randomly assigned to prevent them from stopping.

A total of 10,000 randomly generated episodes (agents with random positions and trajectories) are trained in this study. Each algorithm starts with the same randomly initialized weights to ensure a fair comparison. The training hardware is a computer with an AMD Ryzen 5600X CPU and an Nvidia GeForce RTX 3090 GPU, which can simultaneously train four tasks overall in three days.

5.2. Quantitative evaluation

The baseline of our approach is intrinsic-SGD3QN (Martinez-Baselga et al., 2023), which innovatively introduces intrinsic exploration rewards on top of the related work SG-D3QN (Zhou et al., 2022). Building upon the CrowdNav simulation environment, this work introduces the innovative concept of intrinsic exploration reward. In addition, we incorporate prioritized experience replay, hindsight experience replay, the intrinsic curiosity module with visit count of states, and safety evaluation for exploration. We explore different hyper-parameters and select the best ones in each case. To validate and compare these methods, each method is tested in 10,000 randomly generated episodes in circular scenes. Table 1 compares state-of-the-art methods and our approach, highlighting success rate, collision rate, navigation time, and average return as performance metrics.

The results in the table indicate that our method SafeCrowdNav significantly improves the original results and outperforms other methods. The utilization of prioritized experience replay and hindsight experience replay enhances the efficiency of the agent in utilizing past experiences. Our approach's additional safety evaluation function achieves a success rate of 98.6%, which is a 2% improvement compared to the baseline. Our method also demonstrates the ability to find near-optimal solutions quickly and reduces collision probability by 2%, thereby improving the robustness of navigation.

5.3. Qualitative evaluation

In the simple scenario, the training curve is depicted in Figure 3. The metrics of our method SafeCrowdNav are plotted in orange, AEMCARL (Wang et al., 2022) in blue, Intrinsic-SGD3QN (Martinez-Baselga et al., 2023) in purple and the remaining colors are the metrics of our ablation experiments. It obvious reveals that our method outperforms Intrinsic-SGD3QN (Martinez-Baselga et al., 2023) on four metrics. At the beginning of training, with a randomly initialized model, it is challenging for the agent to accomplish the crowd navigation task, and most of the termination states result in "Timeout" or "Collision." As training progresses, the robot quickly learns to maintain a safe distance from pedestrians. It gradually comprehends the crowd's behavior and plans its path based on its predictions of pedestrian trajectories. The robot's performance becomes relatively stable toward the end of the training.

Through learning-based strategies, the robot is able to reach the target location safely and quickly in both simple and complex scenarios, as depicted in Figures 4A, B. In the complex scenario, the robot needs to pay more attention to avoid pedestrians, resulting in rougher trajectories, and longer navigation times. In both simple and complex scenarios, the robot exhibits proactive, and anticipatory collision avoidance behavior. The robot can recognize and avoid interaction centers where pedestrians approach each

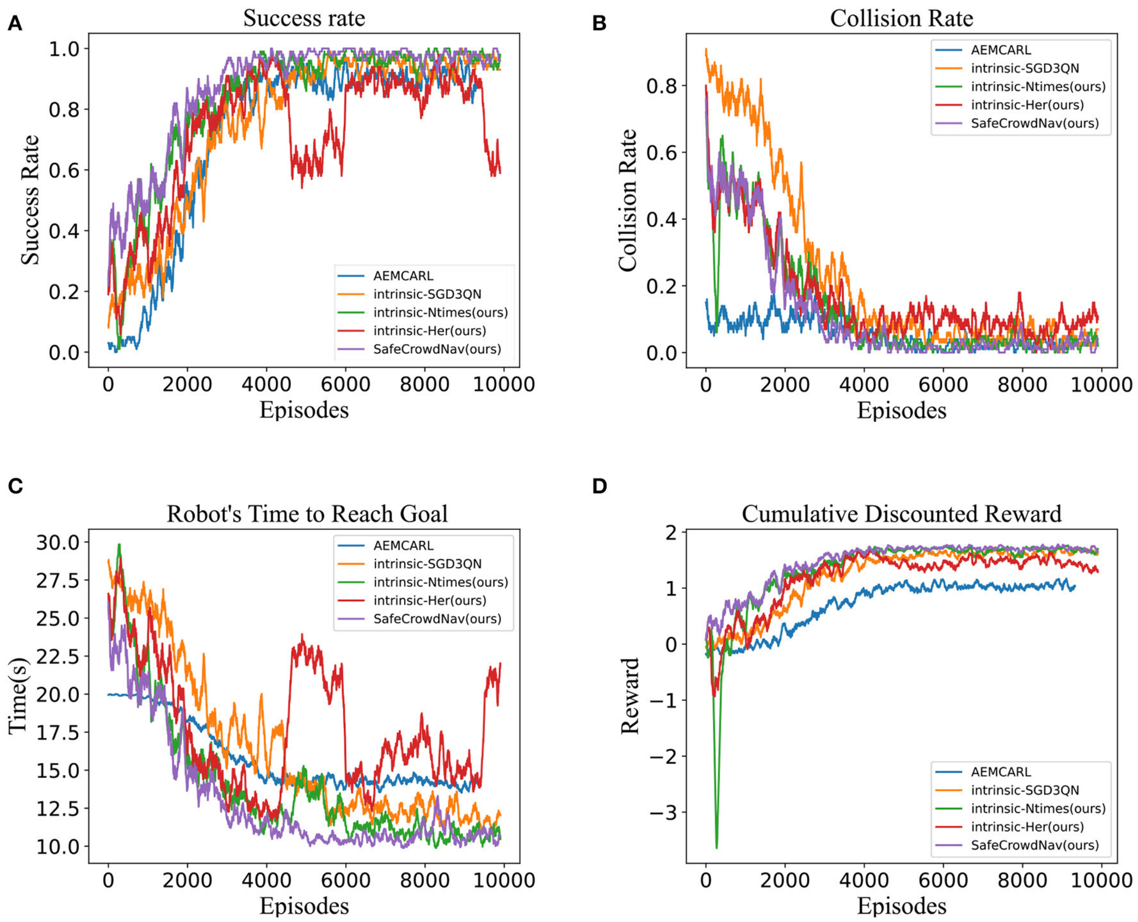


FIGURE 3
Navigation performance about success rate, collision rate, time to reach the goal, and cumulative discounted reward over 10,000 training episodes. **(A)** Success rate. **(B)** Collision rate. **(C)** Time to reach the goal. **(D)** Cumulative discounted reward.

other. For instance, in the simple scenario, the robot suddenly turns right at around 4.0 seconds to avoid a potential encirclement at 5.0 seconds. Additionally, in complex scenarios, even when the robot is surrounded by pedestrians, it possesses the ability to safely escape the environment. In this particular instance, the encirclement by three pedestrians starts at 1.0 seconds and lasts for approximately 3.0 seconds.

The safety evaluation in the tested crowd scenarios is shown in Figure 5, where the real-time safety evaluation score of the robot for the current scene is dynamically displayed. A higher score indicates better safety in the current situation, guiding the robot to navigate faster, while a lower score indicates higher risk, prompting the robot to reduce speed and pay more attention to pedestrians moving toward it or potentially interacting with it. In Figure 5A, the robot's score is 0.46, indicating a lower score due to multiple pedestrians and a complex environment. The lower safety evaluation score guides the robot to reduce speed and allocate different attention weights to surrounding pedestrians, prioritizing obstacle avoidance. In Figure 5B, the robot's score is 0.96, indicating fewer pedestrians in the vicinity and guiding the robot to accelerate its movement, focusing more on navigation tasks. The setting of the safety evaluation score also helps the robot better balance navigation tasks and obstacle avoidance behavior.

6. Conclusion

This paper aims to address safety, autonomy, effectiveness, and user-friendliness in evaluating intelligent robot behaviors. We propose SafeCrowdNav, an innovative approach based on Deep Reinforcement Learning to enhance navigation in crowded environments. Our approach includes heterogeneous spatial-temporal maps for comprehensive environmental representation. We introduce a novel safety evaluation framework based on environment complexity and task difficulty. Additionally, we enhance the intrinsic reward by introducing constraints based on previously encountered scenes, effectively avoiding repetitive and inefficient exploration behavior by the agent. To facilitate efficient and safe navigation in dense crowds, we also integrate prioritized and hindsight experience replay techniques. Extensive evaluations in the CrowdNav simulator demonstrate that SafeCrowdNav achieves

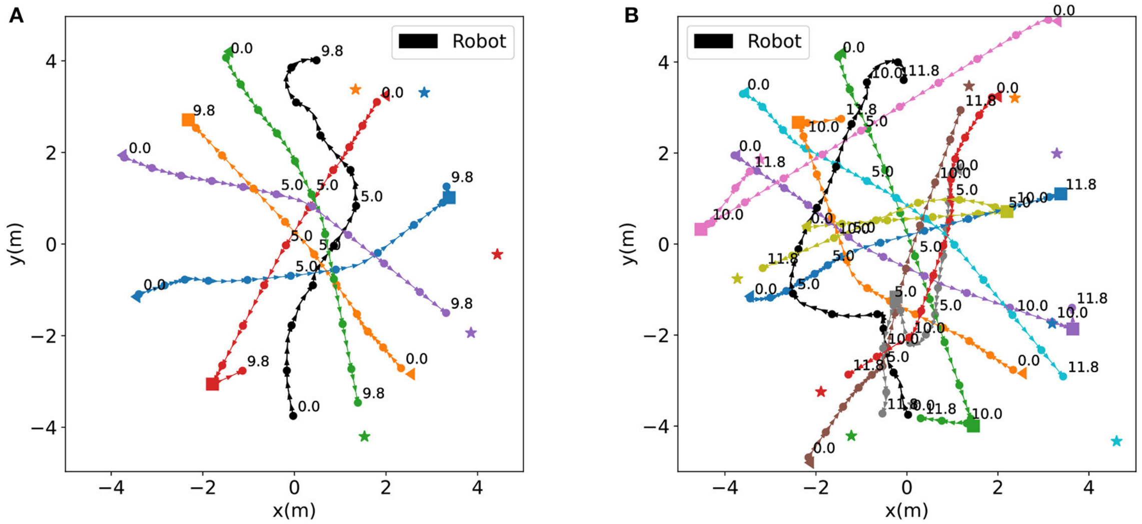


FIGURE 4
Trajectory maps for a simple and a complex scene. In these maps, the circles represent agents, with the black circle representing the robot and other colors representing pedestrians. The numbers near the circles indicate the corresponding time steps. The time interval between two consecutive circles is 1.0 seconds. The maps mark humans' starting positions, turning points, and final goal positions with triangles, squares, and pentagrams, respectively. **(A)** Trajectories in a simple scenario. **(B)** Trajectories in a complex scenario.

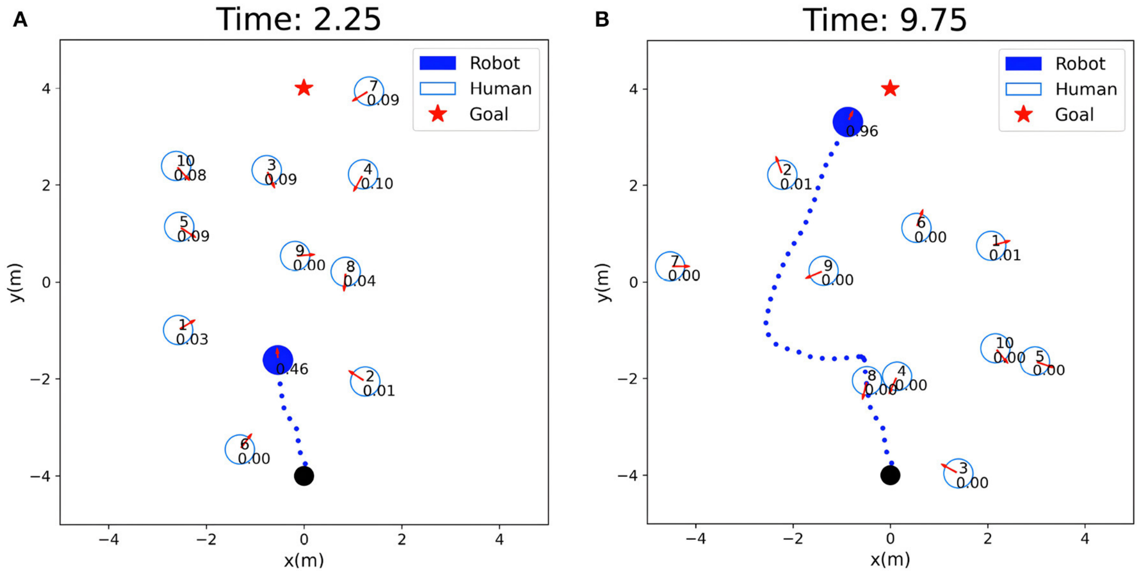


FIGURE 5
Visualization of safety evaluation scores: the solid circle represent the robot, the hollow circles represent humans, and the numbers inside the circles indicate the safety evaluation scores of the robot. **(A)** Low safety evaluation score: 0.46. **(B)** High safety evaluation score: 0.96.

shorter trajectories and higher success rates compared to state-of-the-art algorithms.

However, future works still have many shortcomings to overcome. This includes the need for real-world scenario datasets to enhance performance in real environments, incorporating more realistic human reactions, and exploring the generalization performance from virtual to real-world scenarios. Adjusting the robot's shape based on real-world conditions and conducting real-world observations will provide valuable insights.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

JX: Conceptualization, Formal analysis, Investigation, Methodology, Software, Writing—original draft, Writing—review and editing. WZ: Conceptualization, Writing—original draft. JC: Writing—review and editing, Investigation. HL: Conceptualization, Funding acquisition, Resources, Supervision, Writing—review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work is supported by the National Key R&D Program of China (No. 2020AAA0108904) and the Science and Technology Plan of Shenzhen (No. JCYJ20200109140410340).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., and Savarese, S. (2016). "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 961–971. doi: 10.1109/CVPR.2016.110

Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., et al. (2017). "Hindsight experience replay," in *Advances in Neural Information Processing Systems* 30.

Badia, A. P., Sprechmann, P., Vitvitskyi, A., Guo, D., Piot, B., Kapturowski, S., et al. (2020). Never give up: Learning directed exploration strategies. *arXiv preprint arXiv:2002.06038.*

Cai, J., Huang, W., You, Y., Chen, Z., Ren, B., and Liu, H. (2023). Spsd: Semantics and deep reinforcement learning based motion planning for supermarket robot. *IEICE Trans. Inf. Syst.* 106, 765–772. doi: 10.1587/transinf.2022DLP0057

Chen, C., Hu, S., Nikdel, P., Mori, G., and Savva, M. (2020). "Relational graph learning for crowd navigation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE), 10007–10013. doi: 10.1109/IROS45743.2020.9340705

Chen, C., Liu, Y., Kreiss, S., and Alahi, A. (2019). "Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning," in *2019 International Conference on Robotics and Automation (ICRA)* (IEEE), 6015–6022. doi: 10.1109/ICRA.2019.8794134

Chen, Y. F., Liu, M., Everett, M., and How, J. P. (2017). "Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning," in *2017 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE), 285–292. doi: 10.1109/ICRA.2017.7989037

Duguleana, M., and Mogan, G. (2016). Neural networks based reinforcement learning for mobile robots obstacle avoidance. *Exp. Syst. Applic.* 62, 104–115. doi: 10.1016/j.eswa.2016.06.021

Everett, M., Chen, Y. F., and How, J. P. (2021). Collision avoidance in pedestrian-rich environments with deep reinforcement learning. *IEEE Access* 9, 10357–10377. doi: 10.1109/ACCESS.2021.3050338

Helbing, D., and Molnar, P. (1995). Social force model for pedestrian dynamics. *Phys. Rev. E* 51, 4282. doi: 10.1103/PhysRevE.51.4282

Huang, W., Wan, W., and Liu, H. (2021). Optimization-based online initialization and calibration of monocular visual-inertial odometry considering spatial-temporal constraints. *Sensors* 21, 2673. doi: 10.3390/s21082673

Kästner, L., Buiyan, T., Jiao, L., Le, T. A., Zhao, X., Shen, Z., et al. (2021). "Arena-rosnav: Towards deployment of deep-reinforcement-learning-based obstacle avoidance into conventional autonomous navigation systems," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE), 6456–6463. doi: 10.1109/IROS51168.2021.9636226

Kothari, P., Kreiss, S., and Alahi, A. (2021). Human trajectory forecasting in crowds: A deep learning perspective. *IEEE Trans. Intell. Transp. Syst.* 23, 7386–7400. doi: 10.1109/TITS.2021.3069362

Li, K., Lu, Y., and Meng, M. Q.-H. (2021). "Human-aware robot navigation via reinforcement learning with hindsight experience replay and curriculum learning," in *2021 IEEE International Conference on Robotics and Biomimetics (ROBIO)* (IEEE), 346–351. doi: 10.1109/ROBIO54168.2021.9739519

Liu, H., Qiu, J., and Huang, W. (2022). "Integrating point and line features for visual-inertial initialization," in *2022 International Conference on Robotics and Automation (ICRA)* (IEEE), 9470–9476. doi: 10.1109/ICRA46639.2022.9811641

Liu, S., Chang, P., Huang, Z., Chakraborty, N., Hong, K., Liang, W., et al. (2023). "Intention aware robot crowd navigation with attention-based interaction graph," in *IEEE International Conference on Robotics and Automation (ICRA).* doi: 10.1109/ICRA48891.2023.10160660

Martinez-Baselga, D., Riazuelo, L., and Montano, L. (2023). Improving robot navigation in crowded environments using intrinsic rewards. *arXiv preprint arXiv:2302.06554.* doi: 10.1109/ICRA48891.2023.10160876

Pandey, A., Pandey, S., and Parhi, D. (2017). Mobile robot navigation and obstacle avoidance techniques: a review. *Int. Rob. Auto. J.* 2, 00022. doi: 10.15406/iratj.2017.02.00023

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. (2017). "Curiosity-driven exploration by self-supervised prediction," in *International Conference on Machine Learning* (PMLR), 2778–2787. doi: 10.1109/CVPRW.2017.70

Ray, A., Achiam, J., and Amodei, D. (2019). Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708.*

Sathyamoorthy, A. J., Patel, U., Guan, T., and Manocha, D. (2020). Frozone: Freezing-free, pedestrian-friendly navigation in human crowds. *IEEE Robot. Autom. Lett.* 5, 4352–4359. doi: 10.1109/LRA.2020.2996593

Schaul, T., Quan, J., Antonoglou, I., and Silver, D. (2015). Prioritized experience replay. *arXiv preprint arXiv:1511.05952.*

Trautman, P., and Krause, A. (2010). "Unfreezing the robot: Navigation in dense, interacting crowds," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems* (IEEE), 797–803. doi: 10.1109/IROS.2010.5654369

Van Den Berg, J., Guy, S. J., Lin, M., and Manocha, D. (2011). "Reciprocal n-body collision avoidance," in *Robotics Research: The 14th International Symposium ISRR* (Springer), 3–19. doi: 10.1007/978-3-642-19457-3_1

Van den Berg, J., Lin, M., and Manocha, D. (2008). "Reciprocal velocity obstacles for real-time multi-agent navigation," in *2008 IEEE International Conference on Robotics and Automation* (IEEE), 1928–1935. doi: 10.1109/ROBOT.2008.4543489

Wang, S., Gao, R., Han, R., Chen, S., Li, C., and Hao, Q. (2022). "Adaptive environment modeling based reinforcement learning for collision avoidance in complex scenes," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE), 9011–9018. doi: 10.1109/IROS47612.2022.9982107

Zhou, Z., Zhu, P., Zeng, Z., Xiao, J., Lu, H., and Zhou, Z. (2022). Robot navigation in a crowd by integrating deep reinforcement learning and online planning. *Appl. Intell.* 52, 15600–15616. doi: 10.1007/s10489-022-03191-2

Frontiers in Neurorobotics

TYPE Original Research
PUBLISHED 10 November 2023
DOI 10.3389/fnbot.2023.1288463

OPEN ACCESS

EDITED BY
Long Jin,
Lanzhou University, China

REVIEWED BY
Zhongbo Sun,
Changchun University of Technology, China
Lin Chen,
Chinese Academy of Sciences (CAS), China
Mingchuan Zhou,
Zhejiang University, China

*CORRESPONDENCE
Zhange Liang
✉ liangzhange520@163.com

RECEIVED 04 September 2023
ACCEPTED 30 October 2023
PUBLISHED 10 November 2023

CITATION
Wang M and Liang Z (2023) Cross-modal self-attention mechanism for controlling robot volleyball motion.
Front. Neurorobot. 17:1288463.
doi: 10.3389/fnbot.2023.1288463

Cross-modal self-attention mechanism for controlling robot volleyball motion

Meifang Wang[1] and Zhange Liang[2]*

[1]Sports Department, Anhui Agricultural University, Hefei, China, [2]School of Sports Science, Hefei Normal University, Hefei, Anhui, China

Introduction: The emergence of cross-modal perception and deep learning technologies has had a profound impact on modern robotics. This study focuses on the application of these technologies in the field of robot control, specifically in the context of volleyball tasks. The primary objective is to achieve precise control of robots in volleyball tasks by effectively integrating information from different sensors using a cross-modal self-attention mechanism.

Methods: Our approach involves the utilization of a cross-modal self-attention mechanism to integrate information from various sensors, providing robots with a more comprehensive scene perception in volleyball scenarios. To enhance the diversity and practicality of robot training, we employ Generative Adversarial Networks (GANs) to synthesize realistic volleyball scenarios. Furthermore, we leverage transfer learning to incorporate knowledge from other sports datasets, enriching the process of skill acquisition for robots.

Results: To validate the feasibility of our approach, we conducted experiments where we simulated robot volleyball scenarios using multiple volleyball-related datasets. We measured various quantitative metrics, including accuracy, recall, precision, and F1 score. The experimental results indicate a significant enhancement in the performance of our approach in robot volleyball tasks.

Discussion: The outcomes of this study offer valuable insights into the application of multi-modal perception and deep learning in the field of sports robotics. By effectively integrating information from different sensors and incorporating synthetic data through GANs and transfer learning, our approach demonstrates improved robot performance in volleyball tasks. These findings not only advance the field of robotics but also open up new possibilities for human-robot collaboration in sports and athletic performance improvement. This research paves the way for further exploration of advanced technologies in sports robotics, benefiting both the scientific community and athletes seeking performance enhancement through robotic assistance.

KEYWORDS

multimodal perception, volleyball robot, spiking skill, cross-modal self-attention mechanism, adversarial network, transfer learning

1. Introduction

With the rapid advancement of technology, robotics is gradually permeating various fields, including sports. This study aims to enhance robotic skills in volleyball through deep learning and multimodal sensing technology, injecting innovation, and vitality into the realm of sports (Hong et al., 2021).

High-level sports demand athletes to possess outstanding perceptual, reaction speed, and motor control abilities. The development of modern technology has created opportunities for the application of robotics (Siedentop and Van der Mars, 2022). Robots can serve as ideal practice partners for athletes, enriching the levels and enjoyment of competitions, and offering audiences novel viewing experiences (Siegel and Morris, 2020).

This research is focused on volleyball, a sport characterized by intense teamwork, demanding athletes to make precise decisions and immediate reactions in rapidly changing game scenarios (Oliveira et al., 2020; Weiss et al., 2021). Despite the increasing utilization of robots in sports, there is still room for improvement in robot spiking skills in volleyball. Therefore, this study focuses on improving the skill level of robots in volleyball matches by integrating multimodal perception and deep learning methods, aiming to enable their practical use in real competitions.

In recent years, there has been significant interest and research in the application of robotics technology in the field of sports (Thuruthel et al., 2019; Chen et al., 2020; Oliff et al., 2020). However, despite the extensive research in various sports disciplines, the exploration and study of robotics in volleyball have been relatively limited. Current research primarily focuses on aspects such as robot design, perception, and interaction (Ji et al., 2022; Hu et al., 2023). Nevertheless, there is still a need for further investigation into the application of multimodal perception and deep learning in this context (Olaniyan et al., 2022).

In recent years, there has been a growing interest and research focus on the application of robotics technology in the field of sports (Thuruthel et al., 2019; Chen et al., 2020; Oliff et al., 2020). However, despite extensive research across various disciplines in sports, exploration and research in volleyball robot technology have remained relatively limited. Current research primarily centers around aspects such as robot design, perception, and interaction (Ji et al., 2022; Hu et al., 2023). Jinho So and his colleagues (So et al., 2021) investigated the precise estimation of soft manipulator shape using stretchable shape sensors, while Li and Peng (2022) introduced a monocular visual-tactile sensor to enhance the robustness of robot manipulation. Nevertheless, there is still a need for further research on the application of multimodal sensing and deep learning in this domain (Olaniyan et al., 2022).

The contributions of this paper can be summarized in the following three aspects:

1. This study introduces a cross-modal self-attention mechanism designed to holistically address the amalgamation of diverse multimodal data collected by disparate sensors, including images and action sequences. Leveraging self-attention, we seamlessly integrate information from distinct modalities, enabling robots to comprehensively perceive cyclic motion scenarios. This innovative approach empowers robots to execute various operations with heightened accuracy in repetitive tasks, such as assessing ball velocity, trajectory, and opponent position in volleyball spiking, thus significantly elevating spiking proficiency.
2. The successful application of generative adversarial networks (GANs) to synthesize immersive cyclic motion scenarios is showcased. Through the generative and discriminative processes of GANs, we fabricate authentically textured virtual environments, imbuing robot skill training with heightened challenge and practicality. This augmentation not only fosters skill adaptability but also furnishes an expanded pool of training data, further propelling the prowess of robots.
3. The study maximizes the philosophy of transfer learning, funneling insights gleaned from alternate cyclic motion datasets into the enhancement of robotic skills. This knowledge infusion expedites the robot's mastery of cyclic motion domains, facilitating swift adaptation to competitive settings and accelerated skill growth. This method not only introduces fresh paradigms for robot training but also widens the horizons of transfer learning's applicability in the realm of robotics.

The logical structure of this article is as follows. In Section 2, methods, the technical methods used in this study are introduced in detail, including cross-modal self-attention mechanism, adversarial network, and transfer learning. In Section 3, experiments, the experimental environment and data are described, and the evaluation indicators are introduced. At the same time, the experimental results were analyzed in detail, the performance of different methods and data sets were compared, and the effectiveness of the technical method was verified. In Section 4, discussion and conclusion, the research results are summarized, the significance and contribution of the research are evaluated, the limitations of the research are pointed out, and prospects for future work are proposed.

2. Methodology

In the method part of Chapter 3, we will introduce the overall algorithm flow of this research in detail, and show how to improve the spiking skills of volleyball robots through key technologies such as cross-modal self-attention mechanism, adversarial network, and transfer learning. This comprehensive algorithm process will provide the basis for subsequent experiments and comparative analysis, and also present the overall framework of this study for readers. The overall algorithm flow chart is shown in Figure 1.

2.1. Cross-modal self-attention mechanism

When dealing with multimodal data, attention mechanisms are powerful tools that allow the model to focus on the most relevant information from different modalities. We leverage a cross-modal self-attention mechanism to effectively integrate data from various sensors for enhancing the skills of our volleyball robot (Wang et al., 2021). Attention mechanisms are widely used in deep learning, enabling models to selectively attend to important parts of the data while disregarding irrelevant portions. There are two types of attention mechanisms: self-attention and cross-attention (Niu et al., 2021). Self-attention involves interactions and fusion of information within the same modality. For example, in a language model, self-attention allows each word to adjust its representation based on the context. Cross-attention involves interactions and fusion of information between different modalities. For instance, in visual question-answering tasks, cross-attention can establish correspondences between questions and images. The cross-modal self-attention mechanism is illustrated in Figure 2.

The key to the self-attention mechanism is to calculate the attention weight. One of the classic methods is to use Scaled Dot-Product Attention. Given a set of query vectors (Q), key vectors (K), and value vectors (V), it can compute attention weights

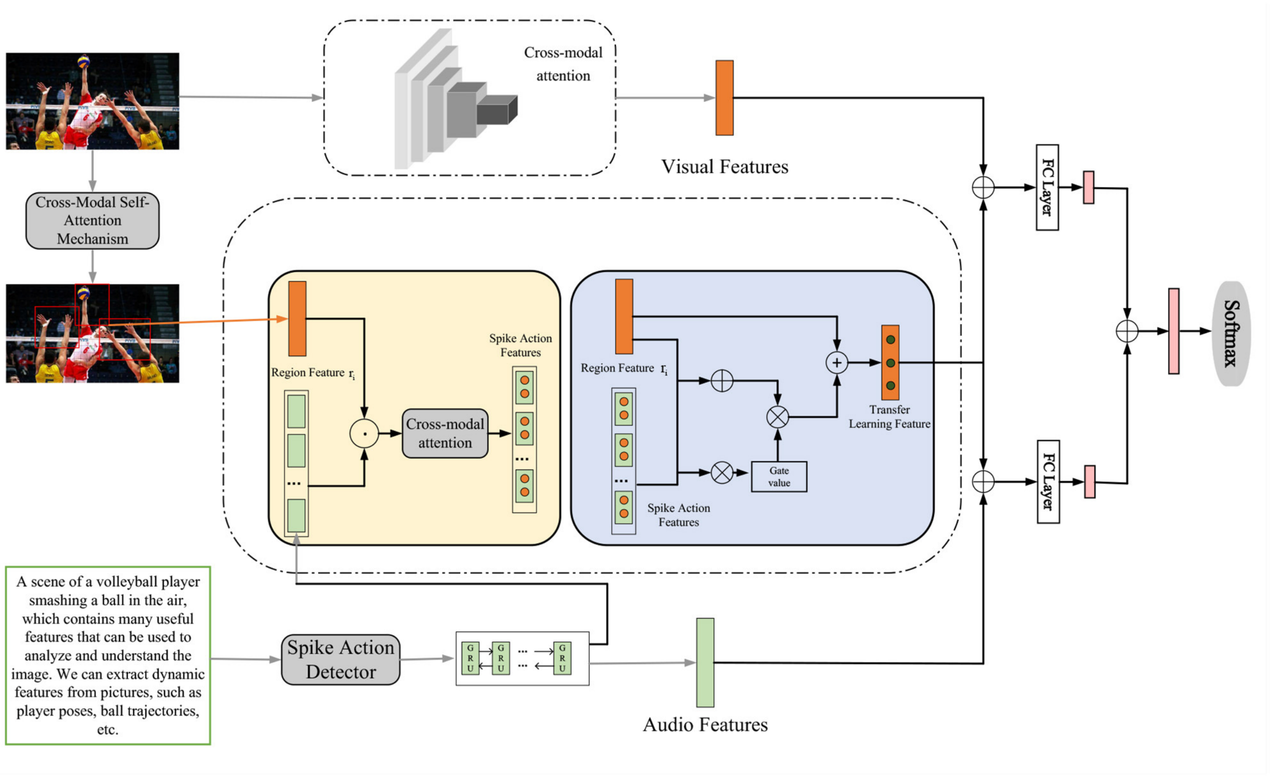

Cross-modal attention
Visual Features
Cross-Modal Self-Attention Mechanism
Region Feature r_i
Spike Action Features
Cross-modal attention
Region Feature r_i
Transfer Learning Feature
Gate value
Spike Action Features
FC Layer
FC Layer
Softmax
A scene of a volleyball player smashing a ball in the air, which contains many useful features that can be used to analyze and understand the image. We can extract dynamic features from pictures, such as player poses, ball trajectories, etc.
Spike Action Detector
GRU
Audio Features

FIGURE 1
Overall algorithm flowchart.

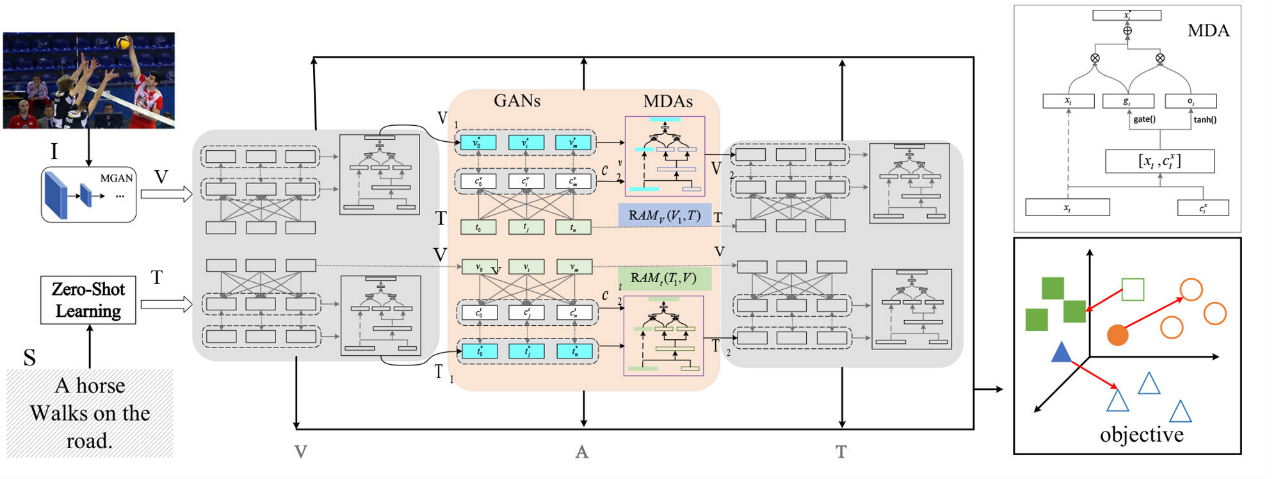


FIGURE 2
Cross-modal self-attention mechanism.

by the following formula:

$$\text{Attention}\,(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{1}$$

where d_k is the dimension of the query and key vectors. The dot product operation in this formula expresses the similarity between the query and the key, and then normalizes using the *softmax* function to get the attention weights. Finally, the weighted values are obtained by multiplying the attention weights with the value vector.

In our study, we further apply the attention mechanism to multimodal data. To synthesize information from different sensors, we introduce a cross-modal self-attention mechanism. In this approach, we take the feature representations of different modalities as queries, keys and values, so that the model can automatically learn the correlation between different modalities.

Formally, suppose we have two modalities M_1 and M_2 with corresponding queries, keys, and values Q_1,K_1,V_1 and Q_2,K_2,V_2, respectively. We can compute the cross-modal self-attention weights as follows:

$$\text{Cross-Modal Attention}\ (Q_1, K_2, V_2) = \text{softmax}\left(\frac{Q_1 K_2^\top}{\sqrt{d_k}}\right) V_2 \tag{2}$$

Similarly, we can calculate the attention weight of modality M_2 to modality M_1.

In practical applications, we also need to consider optimization methods such as loss function and gradient descent to train our model. A commonly used optimization function is the cross-entropy loss function, which has good results in multi-classification tasks. For neural network training, we usually use the backpropagation algorithm to calculate gradients and perform parameter updates. Its formula is as follows:

$$\text{CrossEntropy}\ (p, q) = -\sum_i p_i \log(q_i) \tag{3}$$

where p is the actual probability distribution, q is the probability distribution predicted by the model, and i represents the index of the category. By minimizing the cross-entropy loss, the model can better fit the training data, thus improving the accuracy of predictions.

During the training process of the neural network, we use the backpropagation algorithm to calculate the gradient (Zhang, 2019), and use optimization methods such as gradient descent to update the model parameters. Backpropagation calculates the gradient of each parameter to the loss function through the chain rule, and then uses gradient descent to update the parameters to gradually optimize the model.

Through the cross-modal self-attention mechanism, we can extract key information from different sensor data, realizing the organic fusion and collaboration of multi-modal data. This provides a more solid foundation for our subsequent Generative Adversarial Network and transfer learning. Next, we will detail how to further improve the skills of volleyball robots with the help of Generative Adversarial Network.

2.2. Generative adversarial networks

Generative Adversarial Networks (GANs) are a deep learning framework that consists of two neural networks called a generator and a discriminator (Mi et al., 2020). The goal of a generator is to generate data, such as images, audio, etc., from a random noise vector that has a distribution similar to real data. The goal of the discriminator is to distinguish the data generated by the generator from the real data and give a probability value indicating its authenticity. There is an adversarial relationship between the generator and the discriminator, that is, the generator tries to deceive the discriminator, and the discriminator tries to see through the generator. By alternately training the two networks, the generator is finally able to generate high-quality data, while the discriminator cannot distinguish between real and fake. The confrontation network is shown in Figure 3.

The basic objective function of GAN can be expressed as:

$$\begin{aligned}\min_G \max_D V(D, G) = {} & E_{x\sim p_{data}(x)}[\log D(x)] \\ & + E_{z\sim p_Z(z)}[\log(1 - D(G(z)))]\end{aligned} \tag{4}$$

In this study, we use this method to enhance the spiking skills of a volleyball robot. Our method can effectively utilize the idea of adversarial learning, enabling the generator to learn useful knowledge from other sports game data and transfer it to volleyball games. Our method consists of the following three steps:

1. Data preprocessing. We used a video feature extraction tool to extract the features of each frame in the volleyball video and save it as a feature vector. This tool can use a variety of pre-trained models (such as I3D, I3D-non-local, SlowFast, etc.) to extract powerful video features. We divide each video into segments and label each segment indicating whether the segment contains a smashing action. We regard the clips containing the smashing action as positive samples and the clips not containing the smashing action as negative samples.
2. GANs are trained. We used a Conditional Generative Adversarial Networks to train our model (Xu et al., 2019). Conditional Generative Adversarial Networks is a method of introducing additional information into GANs, such as category labels, text descriptions, etc. The objective function of Conditional Generative Adversarial Networks can be expressed as:

$$\begin{aligned}\text{minmax}_G V(D, G) = {} & E_{x\sim p_{data}(x), y\sim p_{data}(y)}[\log D(x|y)] \\ & + E_{z\sim p_z(z), y\sim p_{data}(y)}[\log(1 - D(G(z|y)))]\end{aligned} \tag{5}$$

Among them, $V(D, G)$ is the objective function of GANs,$D(x)$ is the probability that the discriminator gives the input x is real data, $G(z)$ is the data generated by the generator from the noise vector z, $p_{data}(x)$ is the real data distribution, and $p_z(z)$ is the noise vector distribution. y is extra information, such as category labels. In our method, we use a textual description as additional information, indicating the requirement of the smashing action, such as "the smashing angle is 45 degrees, and the force is 80%". The training process of GANs can be regarded as a zero-sum game, that is, the discriminator and the generator compete with each other so that the objective function reaches the Nash equilibrium, namely:

$$G^* = \underset{G}{\text{argminmax}} V(D, G) \tag{6}$$

3. GANs application. We use a decoder to restore the sequence of feature vectors of the video clips produced by the generator to a sequence of images, which are stitched into a single video. We compare the generated videos with real volleyball match videos to evaluate their quality and authenticity. We also use the generated videos as training data for the volleyball robot to improve its spiking skills. The output of the decoder can be expressed as:

$$\hat{x}_t = f_{dec}(h_t) \tag{7}$$

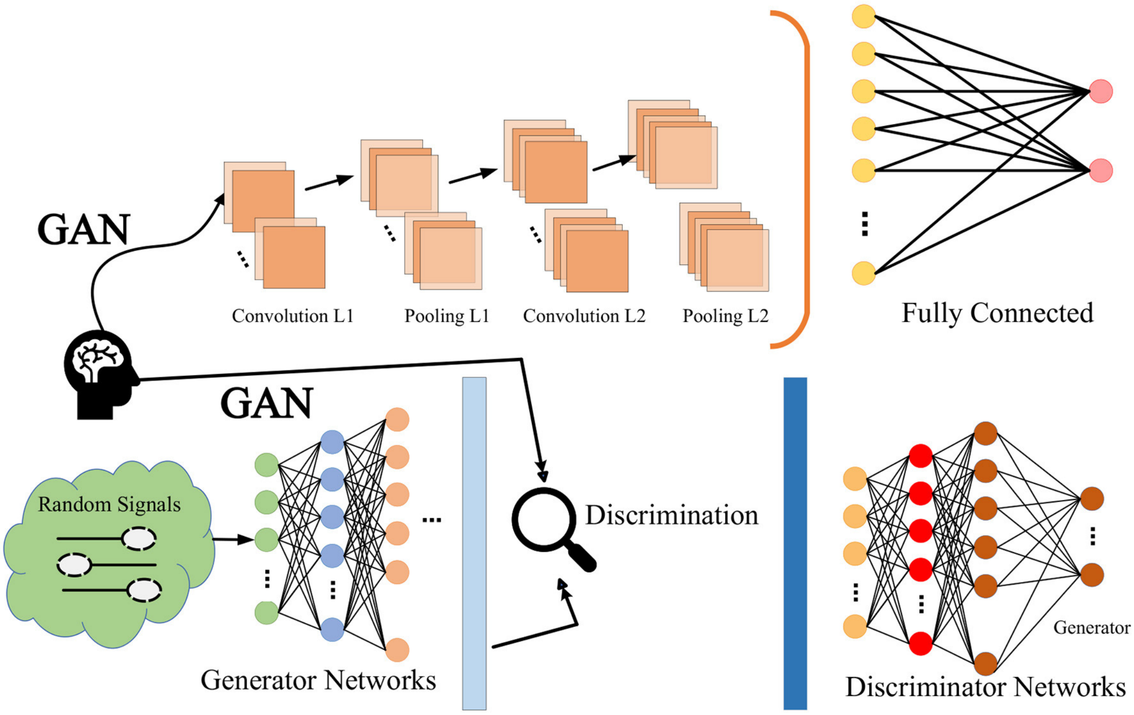


FIGURE 3
Generative adversarial networks.

Among them, $\hat{x}_t$ is the image generated at the t-th moment, f_{dec} is the decoder function, and h_t is the feature vector output by the generator at the t-th moment.

During training, the generator and discriminator are optimized through adversarial learning, specifically, the generator tries to minimize $V(D, G)$, while the discriminator tries to maximize $V(D, G)$. This leads to a dynamic balancing process at which the samples generated by the generator are realistic enough that the discriminator cannot effectively distinguish real samples from generated samples. In terms of optimization functions, for the generator G, we can use the following optimization functions to update the parameters of the generator:

$$\min_G V(D, G) = \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{8}$$

In practical applications, through methods such as backpropagation and gradient descent, the parameters of the generator and discriminator can be gradually optimized to achieve the goal of training GANs.

By introducing GANs, we can further improve the skills of volleyball robots and generate more realistic and diverse game scenes, thus laying a more solid foundation for the improvement of robot skills. Next, we explore how transfer learning can be applied to skill improvement for volleyball robots.

2.3. Transfer learning

We use a transfer learning method called domain adaptation (Zhuang et al., 2020). In this approach, we improve the generalization of the model on the target domain by minimizing the domain difference between the source and target domains. Transfer learning is shown in the Figure 4.

Assuming we have source domain data D_{souree} and target domain data D_{target}, our goal is to transfer the knowledge on the source domain to the target domain. We can achieve this by minimizing the distribution difference between the source and target domains. A common method is Maximum Mean Difference (MMD):

$$\text{MMD}\left(\mathcal{D}_{\text{source}}, \mathcal{D}_{\text{target}}\right) = \left\| \frac{1}{n_{\text{source}}} \sum_{i=1}^{n_{\text{source}}} \phi(x^i_{\text{source}}) - \frac{1}{n_{\text{target}}} \sum_{j=1}^{n_{\text{target}}} \phi(x^j_{\text{target}}) \right\|^2 \tag{9}$$

Among them, x^i_{source} and x^i_{target} denote samples in the source domain and target domain, respectively, and $\phi(.)$ is a mapping function that maps samples into a latent space. By minimizing MMD, we can reduce the distribution difference between source and target domains, thus enabling transfer learning.

Another common approach is Domain Adversarial Neural Network (DANN) (Ajakan et al., 2014). In DANN, we introduce a domain classifier whose goal is to distinguish samples in the source and target domains. At the same time, we train a feature extractor to generate features that are indistinguishable to domain classifiers. This can be achieved by minimizing the loss function of the domain classifier:

$$\mathcal{L}_{\text{domain}} = -\frac{1}{n} \sum_{i=1}^{n} \log D(f(x_i)) \tag{10}$$

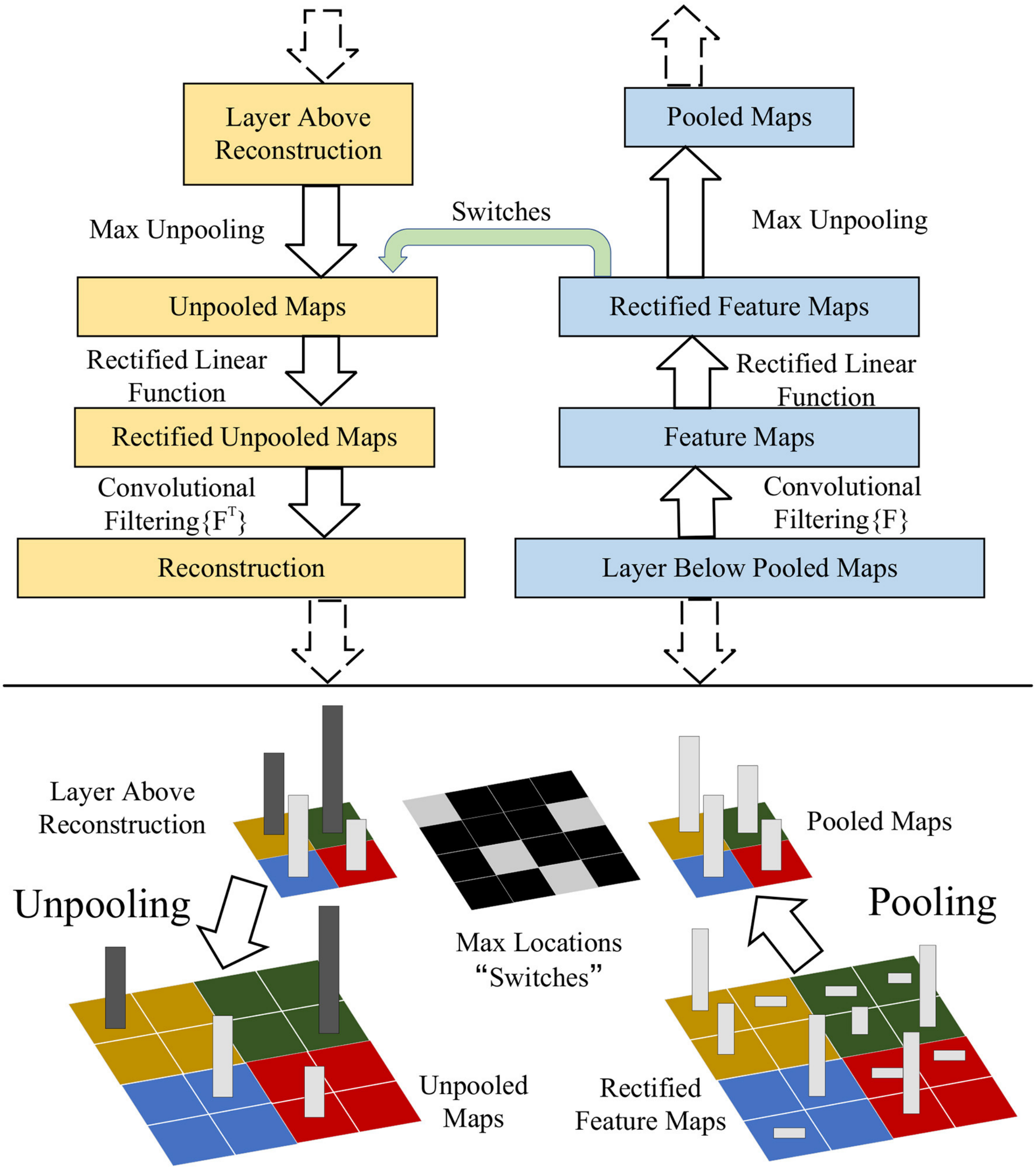


FIGURE 4
Interaction with key modules through transfer learning, including feature extractors, self-attention mechanisms, and robot controllers. These modules were optimized by transfer learning to achieve better performance.

Among them, $D(.)$ is the domain classifier, $f(.)$ is the feature extractor, and E is the sample. By minimizing L_{domain}, we can make features more consistent across domains, enabling transfer learning.

In addition, there is a common method of transfer learning by training an initial model on the source domain, then using the parameters of this model as the initial parameters of the target domain model, and then fine-tuning the model parameters on the target domain. This can be achieved by minimizing a loss function over the target domain:

$$\mathcal{L}_{\text{target}}(f_{\text{target}}, \mathcal{D}_{\text{target}}) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{target}}}\left[\ell(f_{\text{target}}(x), y)\right] \tag{11}$$

Among them, f_{target} is the model on the target domain, D_{target} is the data distribution of the target domain, (x, y) is the sample of the target domain, and ℓ represents the loss function.

Optimization methods for transfer learning usually consist of two steps: feature extraction and fine-tuning. In the feature extraction stage, we can extract general feature representations from the source domain through pre-trained models. Then, in the fine-tuning stage, we train the extracted feature representations together with data from the target domain to further adapt to the target domain. Specifically, the fine-tuning optimization function can be expressed as:

$$\mathcal{L}_{\text{target}}(f_{\text{target}}, D_{\text{target}}) + \lambda \cdot \mathcal{L}_{\text{source}}(f_{\text{target}}, D_{\text{source}}) \quad (12)$$

Among them, L_{source} represents the loss function on the source domain, and λ is a hyperparameter that weighs the two losses.

Through transfer learning, we can make full use of the knowledge of the existing modality in the task of volleyball robot and improve the performance of the model in the new modality.

In the Section 2 of this chapter, we propose a method that comprehensively applies attention mechanisms, GANs, and transfer learning to improve the skills of volleyball robots. First, we introduce a cross-modal self-attention mechanism, which effectively integrates multi-modal sensor data, enabling the model to automatically learn the correlation between different modalities. By calculating attention weights, we are able to extract key information from different sensor data, laying a solid foundation for the subsequent steps. Then, we introduced the application of GAN. Through domain adaptation and domain confrontation neural network, the knowledge transfer between the source domain and the target domain is realized, thereby improving the generalization ability of the model in the target domain. Finally, we explore how to train the initial model on the source domain and fine-tune the parameters on the target domain to fit the data distribution of the target domain through transfer learning. The comprehensive application of these methods provides strong support for our experimental part. In the next chapter, we will introduce the experimental design and result analysis in detail to verify the effectiveness and performance improvement of our proposed method in improving the skills of volleyball robots.

3. Experiment

The experimental process of this paper is shown in Figure 5.

3.1. Experimental environment

- Hardware Environment

  In this research, we rely on an advanced computing platform as the hardware environment, which is equipped with a high-performance AMD Ryzen 7 5800X processor, equipped with 64GB ultra-high-speed DDR4 memory, and configured with 2 NVIDIA GeForce RTX 3080 10GB graphics card. This excellent hardware configuration endows us with powerful computing and storage capabilities, especially suitable for training and inference of deep learning tasks. In addition, we also use multi-channel SSD hard disk to ensure the high efficiency of data reading and storage. Such a hardware environment provides strong support for the smooth progress of the experiment, making the training process of the model more efficient, stable, and reliable.
- Software Environment

  In this study, we used Python and PyTorch to implement a method for improving the spiking skills of volleyball robots based on deep learning. As the main deep learning framework, PyTorch provides us with powerful model building and training tools, allowing us to flexibly design and optimize our spiking skill model. In the experiment, we made full use of PyTorch's efficient computing power and automatic differentiation function to speed up the model training process, so that our model can converge faster and achieve better results.

3.2. Experimental data

- Volleyball Dataset

  Volleyball Dataset is a video action recognition dataset proposed by Ibrahim et al. of Simon Fraser University in Canada in 2016. The data set consists of 55 volleyball game videos, in which 4830 key frames mark the player's position, individual action and group behavior. Single action includes 9 categories, such as smash, block, pass, etc. Group behavior includes 8 categories, such as passing the ball to the left, scoring from the right, and both sides scrimmage. This dataset aims to provide a challenging scenario for studying the recognition and understanding of human actions and group activities in videos. It can be used for a variety of video analysis tasks, such as action recognition, group activity recognition, person tracking, etc. This dataset has been used and cited by several research papers, demonstrating its value and influence in the field of video analysis.
- VREN: Volleyball Rally Dataset with Expression Notation Language

  VREN is a video volleyball game dataset proposed by Xia et al. at the University of California, Santa Barbara in 2022. This dataset contains video clips from professional and NCAA Div-I indoor volleyball matches, where each round is annotated with a volleyball description language. This language can completely describe the player's action, position, and volleyball trajectory in the volleyball game. This dataset aims to provide a rich and high-level benchmark for studying the skills of robots in volleyball games. Based on the language, this dataset proposes three tasks for automated volleyball action and tactical analysis: (1) volleyball round prediction, which aims to predict the outcome of rounds and help players and coaches improve decision-making in practice; (2) setter type and Smash type prediction, helping athletes, and coaches to prepare for the game more effectively; (3) Volleyball tactics and offensive zone statistics, providing advanced volleyball statistics to help coaches better understand the game and opponent's tactics. The authors conduct a case study showing how experimental results can provide insight to the volleyball analysis community. Furthermore, experimental evaluations on real data establish a baseline for future research and applications. The research bridges

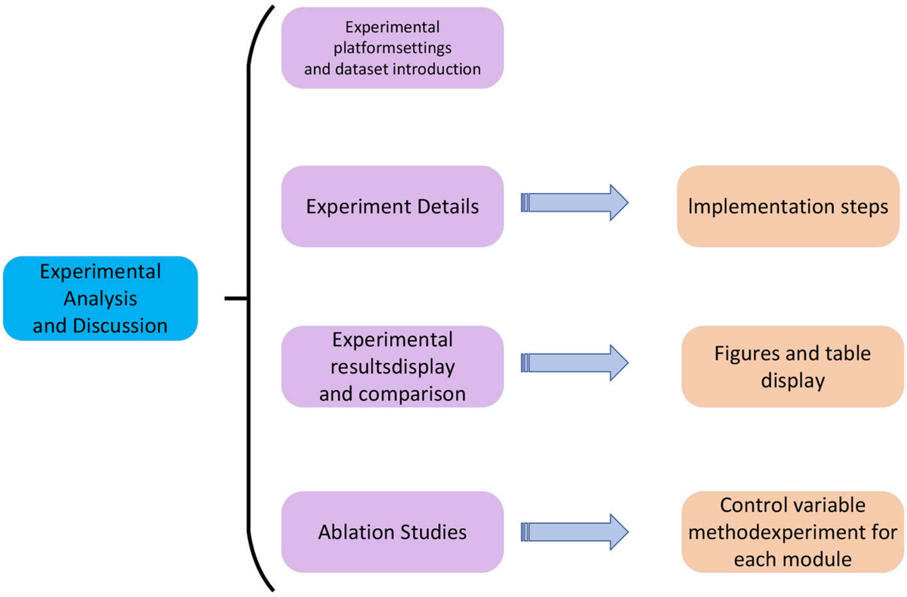

Experimental Analysis and Discussion
Experimental platformsettings and dataset introduction
Experiment Details
Implementation steps
Experimental resultsdisplay and comparison
Figures and table display
Ablation Studies
Control variable methodexperiment for each module

FIGURE 5
The flow chart of the experiment.

the gap between the field of indoor volleyball and computer science.

- UCF101

  The UCF101 dataset is a video action recognition dataset proposed by Soomro et al. at the University of Central Florida in 2012. The dataset consists of 13,320 real action videos from YouTube, covering 101 action categories. These action categories can be divided into five types: human-object interaction, body movement, human-human interaction, playing musical instruments, and sports, some of which are related to volleyball, such as smashing, blocking, passing, etc. This dataset is an extension of the UCF50 dataset, which has only 50 action categories. The UCF101 dataset is highly diverse and challenging because there are a large number of changes in camera motion, object appearance and pose, object scale, viewing angle, background clutter, and lighting conditions in the video. This dataset aims to facilitate further research in the field of action recognition by learning and exploring new categories of real actions.
- MultiSports dataset

  The MultiSports dataset is a video multiplayer sports action detection dataset, which was proposed by Li et al. of Nanjing University in 2021. The dataset consists of 3200 video clips of sports games from YouTube, covering 4 sports categories: aerobics, basketball, football, and volleyball. The dataset annotates 37,701 action instances and 902k bounding boxes, and each action instance has a fine-grained action category label, such as smashing, blocking, passing, etc. This dataset aims to provide a rich and challenging benchmark for studying multi-person video action detection. The dataset has the characteristics of high diversity, high density, and high quality, and can reflect real sports competition scenes.

3.3. Evaluation index

In the assessment process of this research, in order to comprehensively and objectively measure the effectiveness and performance of the proposed sports teaching method, a series of key evaluation metrics were employed. These metrics not only facilitate a quantitative evaluation of the model's performance across various tasks but also provide us with in-depth insights to better comprehend the strengths and limitations of the method. In the following section, we will provide a detailed introduction and analysis of the following key metrics: accuracy, recall, precision, and F1 score. These metrics will assist us in objectively evaluating the efficacy of the proposed method in the context of sports teaching, thereby providing robust support for the reliability of the research and the feasibility of its practical application.

- Hit rate

  In the skill improvement task of the volleyball robot, the hitting rate is a critical evaluation metric used to measure the performance of the proposed method. The hitting rate is defined as the ratio between the number of events correctly predicted by the model and the total number of samples. It provides an intuitive reflection of the model's prediction accuracy, aiding in the assessment of its performance. The hitting rate can be calculated using the following formula:

$$\text{Hit Rate} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{13}$$

  In the context of skill enhancement tasks for the volleyball robot, the hitting rate is a pivotal evaluation metric used to gauge the performance of the proposed method. The hitting rate is defined as the ratio between the number of correctly predicted positive samples (True Positives, TP), indicating the number of instances where skill improvement was accurately identified, and the total number of samples. Similarly, the correctly predicted negative samples (True Negatives, TN) represent instances where the absence of skill improvement was accurately recognized. Conversely, the false positives (FP) correspond to instances where the model erroneously predicted positive samples when they were, in fact, negative. The false negatives (FN) denote cases where the model inaccurately predicted negative samples as positive.

By calculating the hitting rate, we gain insights into the model's accuracy in predicting skill levels, thereby evaluating the practicality and effectiveness of this approach in real-world sports teaching scenarios. In our research, the hitting rate will serve as a critical evaluation metric, assisting us in conducting a thorough analysis of model performance and providing robust support for subsequent experimental findings.

- Recall

In the skill enhancement task of the volleyball robot, recall is a critical evaluation metric used to assess the effectiveness of the proposed attention-based mechanism in capturing the skill level of volleyball players. Recall measures the model's ability to correctly identify actual positive samples, i.e., the proportion of samples that the model correctly predicts out of all actual positive samples. This is of significant importance for evaluating the model's overall performance in sports education. Recall can be calculated using the following formula:

$$\text{Recall} = \frac{TP}{TP + FN} \times 100\% \tag{14}$$

In the context of skill enhancement tasks for the volleyball robot, recall is a crucial evaluation metric used to assess the model's ability to correctly identify positive samples. Specifically, it measures the proportion of samples that the model accurately predicts as skill level improvements out of all actual positive samples in the volleyball skill enhancement task. Conversely, false negatives (FN) represent the positive samples that the model fails to predict accurately, indicating instances where skill level improvements were missed.

By computing the recall rate, we gain insights into the model's capacity to recognize positive samples, thus evaluating the effectiveness of the attention-based mechanism in enhancing the volleyball robot's skills. In our research, recall will serve as a vital evaluation metric, enabling us to conduct an in-depth analysis of model performance, providing a comprehensive assessment, and supporting subsequent experimental results.

- Precision

In the context of skill enhancement tasks for the volleyball robot, precision is a critical evaluation metric used to measure the accuracy of the attention-based method in predicting the skill level of volleyball players. Precision assesses the proportion of samples that the model predicts as positive samples, which are indeed positive samples in reality. This is of paramount importance for evaluating the reliability and accuracy of the model in sports education. Precision can be calculated using the following formula:

$$\text{Precision} = \frac{TP}{TP + FP} \times 100\% \tag{15}$$

TP (True Positives): The number of positive samples correctly predicted by the model, indicating the instances where skill level improvement was accurately identified. FP (False Positives): The number of positive samples incorrectly predicted by the model, signifying the instances where the model erroneously predicted negative samples as positive.

By calculating precision, we gain insights into the model's accuracy when predicting positive samples, thereby evaluating the effectiveness of the attention-based method in the volleyball robot's skill enhancement task. In our research, precision will serve as a crucial evaluation metric, aiding us in analyzing model performance, providing a dependable assessment, and supporting our experimental results.

- F1 Score

In our study of skill enhancement in volleyball robots, the F1 score serves as a critical evaluation metric employed for the comprehensive assessment of the method's performance in skill improvement. This score takes into account both precision and recall, thus facilitating the equilibrium between the model's accuracy and comprehensiveness in identifying skill improvement instances. Consequently, it provides a more comprehensive performance measurement metric. The formula for calculating the F1 score is as follows:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\% \tag{16}$$

In this formula, we introduce previously discussed precision and recall as parameters. Precision measures the accuracy of the model in identifying positive samples as positive samples, while recall gauges the model's comprehensive recognition capability of positive samples.

The F1 score combines the accuracy and comprehensiveness of the model in skill improvement cases, making it a crucial evaluation metric in the volleyball skill enhancement study. By calculating the F1 score, we can gain a more comprehensive understanding of the method's performance, ensuring that the model achieves accurate and comprehensive results in skill improvement.

Algorithm 1 represents the algorithm flow of the training in this paper.

3.4. Experimental comparison and analysis

In the preceding sections, we provided a comprehensive introduction to the design and implementation of the multimodal perception-based deep learning approach for enhancing volleyball robot spiking skills. In this chapter, our focus shifts toward a comparative analysis of experimental results, aiming to comprehensively evaluate the effectiveness and superiority of the proposed methods. By conducting experiments on multiple datasets, our goal is to delve into the contributions of different

```
Input: Volleyball Dataset, VREN Dataset, UCF101
Dataset, MultiSports Dataset
Initialize Trans-GAN Net parameters: $\theta$
Initialize discriminator parameters: $D$
Define cross-modal self-attention mechanism:
$Attention(X)$
for each epoch do
  for each dataset in [Volleyball, VREN, UCF101,
  MultiSports] do
    Load batch of data: $X$
    Compute attention weights: $W = Attention(X)$
    Compute transformed features: $X' = W \times X$
    Generate fake data: $X_{fake} = TransGAN(X')$
    Compute discriminator loss: $L_D = -\log(D(X_{fake})) - \log(1 - D(X))$
    Compute generator loss: $L_G = -\log(D(X_{fake}))$
    Backpropagate and update $\theta$ and $D$ using $L_D$
    and $L_G$
  end for
  if epoch % transfer interval == 0 then
    Perform transfer learning by copying
    features to next layers
  end if
  Compute evaluation metrics on validation set:
  Hit Rate: $\frac{\sum_{i=1}^{N} 1(y_i = \hat{y}_i)}{N}$
  Recall: $\frac{\sum_{i=1}^{N} 1(y_i = \hat{y}_i \text{ and } y_i = 1)}{\sum_{i=1}^{N} 1(y_i = 1)}$
  Precision: $\frac{\sum_{i=1}^{N} 1(y_i = \hat{y}_i \text{ and } y_i = 1)}{\sum_{i=1}^{N} 1(\hat{y}_i = 1)}$
  F1 Score: $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$
  if F1 Score > best score then
    Save best model
  end if
end for
```

Algorithm 1. Training of Trans-GAN Net.

models and their combinations in enhancing volleyball robot skills, as well as to validate the applicability of our approach across various scenarios. This process of experimental comparison and analysis not only directly showcases the practical effectiveness of our approach but also provides deeper insights, guiding us toward optimizing and advancing the technological trajectory of sports robots.

Through comparing experimental results across different datasets, we will uncover the performance of the multimodal perception-based deep learning approach in varying contexts. Simultaneously, we will integrate the evaluation metrics introduced earlier, such as hit rate, recall rate, precision, and F1 score, to conduct a comprehensive assessment of the overall model performance. We will also analyze the introduction of different modules, exploring the specific roles of cross-modal self-attention mechanisms, GANs, transfer learning, and other methods in enhancing volleyball robot skills. In-depth analysis of the experimental results will allow us to understand the interplay between different modules and their impact on enhancing robot skills.

Furthermore, we will compare the experimental results with those of the baseline models to quantify and illustrate the superiority of our approach. Through comparative analysis, we can accurately evaluate the performance improvement brought about by the multimodal perception-based deep learning approach in enhancing volleyball robot skills. These comparative and analytical results will further validate the feasibility and practicality of our approach, providing robust support and references for research and applications in the field of sports robotics.

Next, we will meticulously dissect the experimental results, comprehensively showcasing the performance of our model across different datasets and metrics, providing readers with a comprehensive understanding of the model's capabilities and its potential value in real-world applications.

From the data in Table 1 above, it can be seen that our method outperforms other research works on both the Volleyball dataset and the VREN dataset. Specifically, on the Volleyvall data set, after removing our method, compared with the research method of Salim et al., who achieved the highest hit rate of 91.66% and the F1 score of 90.77%, our hit rate It has increased by 4.45%, and the F1 score has also increased by 3.98%. At the same time, our precision and recall rate have also reached the optimal value of all methods, reaching 95.41 and 94.57%, respectively, and the performance on the VREN data set is also better than other methods, our hit rate and The F1 score is 8.33 and 6.31% higher than the research method of Kautz et al., and 7.03 and 4.28% higher than the method of Liang et al. In general, from the evaluation results on these two classic volleyball datasets, it can be seen that our new deep learning method with multi-modal information learning and deep generative network as the backbone is effective in identifying and predicting volleyball. There is a significant advantage in action. It can better learn and mine the visual and motion features in volleyball, so it has higher precision and recall. This shows that the method has great potential in improving the motion control skills of volleyball robots. Finally, we compared and visualized the results in Table 1, as shown in Figure 6.

According to the comparative data of Hit Rate, Recall, Precision, and F1 Score of different methods on the two datasets in Table 2 above, it can be seen that our method has significant advantages over other methods. On UCF101 dataset and MultiSports dataset, compared with the work of Kautz et al. using the same dataset, our proposed method achieves 9.57% higher hit rate and 7.79% higher recall rate on UCF101 dataset, F1 The score is 8.48% higher; the hit rate is 7.85% higher in the MultiSports dataset, the recall rate is 6.78% higher, and the F1 score is 7.46% higher. At the same time, excluding our method, compared with Salim et al.'s study on UCF101 which obtained the highest recall rate of 90.81%, our recall rate improved by 3.86%. Compared with Tang et al. who obtained F1 score of 88.20% in the MultiSports dataset, our F1 score increased by 6.88%. Furthermore, we exceed the main evaluation metrics of other methods such as Liang et al. and Wenninger et al. on these two action datasets. This shows that the method shows stronger generalization ability in learning joint motion and action features, and can better identify and classify different types of sports actions. Overall, its excellent performance on two large-scale general-purpose motion datasets once again confirms the advantages of this method in the field of action recognition. We compared and visualized the results in Table 2, as shown in Figure 7.

TABLE 1 Comparison of Hit Rate, Recall, Precision, and F1 Score indicators based on different methods under Volleyball and VREN datasets.

Model	Datasets							
	Volleyball dataset (Ibrahim et al., 2016)				VREN dataset (Xia et al., 2022)			
	Hit rate (%)	Recall (%)	Precision (%)	F1 Score (%)	Hit rate (%)	Recall (%)	Precision (%)	F1 Score (%)
Kautz et al. (2017)	87.28	87.41	88.87	88.13	86.69	87.23	88.37	87.8
Li and Tian (2023)	88.24	87.75	87.93	87.84	86.54	88.73	88.15	88.44
Tang (2021)	89.02	88.88	88.47	88.67	87.82	89.74	89.46	89.6
Liang and Liang (2022)	89.47	89.57	88.59	89.08	87.99	89.79	89.88	89.83
Wenninger et al. (2020)	89.98	90.68	88.96	89.81	89.81	89.89	91.76	90.82
Salim et al. (2019)	91.66	90.95	90.59	90.77	90.02	89.99	92.31	91.14
Ours	96.11	95.41	94.57	94.99	95.02	92.02	96.29	94.11

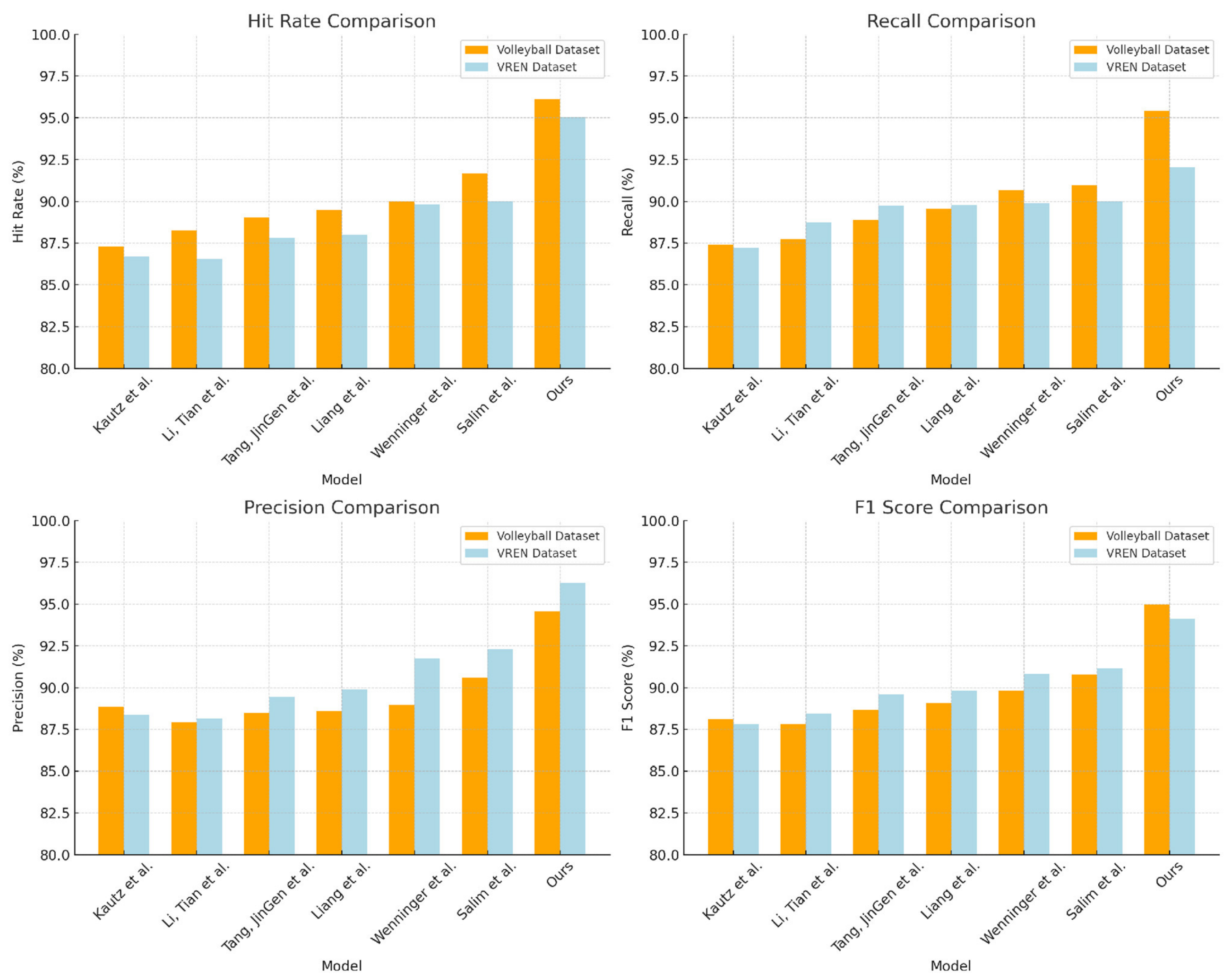


FIGURE 6
Comparison and visualization of Hit Rate, Recall, Precision, and F1 Score indicators based on different methods under Volleyball and VREN datasets.

According to the data in Table 3 above, with the improvement of the model structure, the performance of our proposed method on the two classic volleyball data sets has been significantly improved. Specifically, compared with the baseline model, after adding the self-attention mechanism, the hit rate on the Volleyball dataset increased by 6.78%, the recall rate increased by 11.76%,

TABLE 2 Comparison of Hit Rate, Recall, Precision, and F1 Score indicators based on different methods under UCF101 and MultiSports datasets.

Model	Datasets							
	UCF101 dataset (Soomro et al., 2012)				MultiSports dataset (Li et al., 2021)			
	Hit rate (%)	Recall (%)	Precision (%)	F1 Score (%)	Hit rate (%)	Recall (%)	Precision (%)	F1 Score (%)
Kautz et al. (2017)	86.71	86.88	87.84	87.36	88.83	86.75	88.51	87.62
Li and Tian (2023)	85.87	87.32	88.11	87.71	89.3	86.67	87.98	87.32
Tang (2021)	86.89	88.41	89.35	88.88	90.49	87.56	88.84	88.2
Liang and Liang (2022)	87.36	89.69	90.3	89.99	91.19	88.83	90.05	89.44
Wenninger et al. (2020)	89.54	90.29	91.04	90.66	91.74	89.7	91.27	90.02
Salim et al. (2019)	90.6	90.81	91.58	91.19	92.06	90.22	91.34	90.78
Ours	96.28	94.67	97.03	95.84	96.68	93.53	96.68	95.08

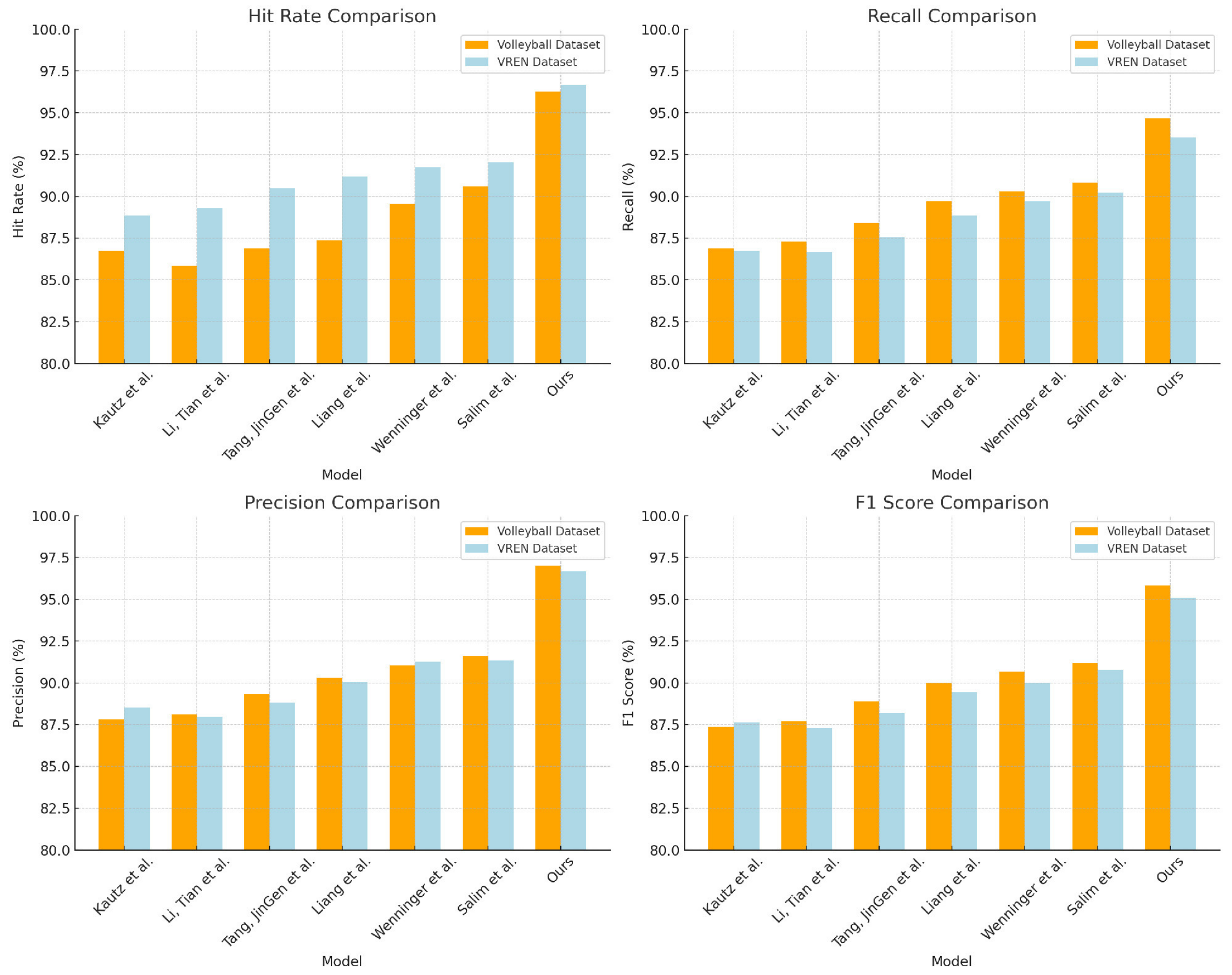


FIGURE 7
Comparison and visualization of Hit Rate, Recall, Precision, and F1 Score indicators based on different methods under UCF101 and MultiSports datasets.

and the F1 score increased by 7.15%; the corresponding increase in the VREN dataset They are 7.49, 5.27, and 6.41%, respectively. After adding the generative adversarial network to the attention model, the indicators of the two data sets have been further improved. Among them, the hit rate and F1 score of the Volleyball data set have increased by about 9.73 and 7.5%,

TABLE 3 Comparison and visualization of Hit Rate, Recall, Precision, and F1 Score indicators of different modules based on Volleyball and VREN datasets.

Module	Dataset							
	Volleyball dataset (Ibrahim et al., 2016)				VREN dataset (Xia et al., 2022)			
	Hit rate (%)	Recall (%)	Precision (%)	F1 Score (%)	Hit rate (%)	Recall (%)	Precision (%)	F1 Score (%)
baseline	65.49	64.34	67.57	65.92	66.52	67.09	67.39	67.24
+satt	72.27	76.10	70.28	73.07	74.01	72.36	74.99	73.65
+gan	82.00	78.58	82.67	80.57	80.12	84.36	77.02	80.52
+satt gan(our)	95.81	93.72	95.71	94.70	96.15	94.85	96.15	95.49

"satt" is the self-attention mechanism, and "gan" is the generative adversarial network.

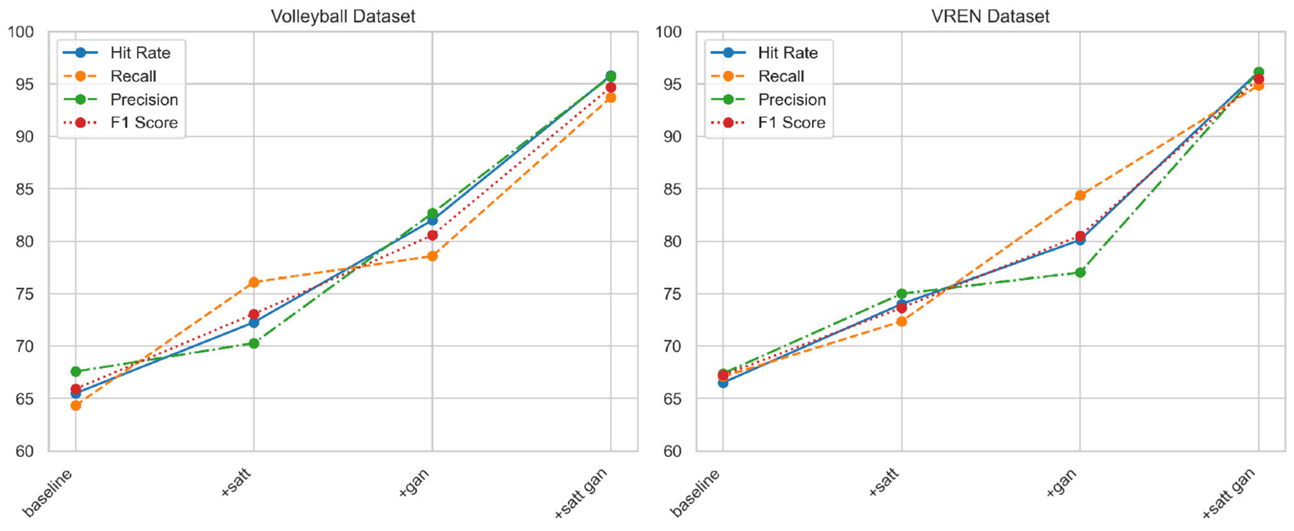


FIGURE 8
Comparison and visualization of Hit Rate, Recall, Precision, and F1 Score indicators of different modules based on Volleyball and VREN datasets.

TABLE 4 Comparison of Hit Rate, Recall, Precision, and F1 Score indicators of different modules based on UCF101 and MultiSports datasets.

Module	Dataset							
	UCF101 dataset (Soomro et al., 2012)				MultiSports dataset (Li et al., 2021)			
	Hit rate (%)	Recall (%)	Precision (%)	F1 Score (%)	Hit rate (%)	Recall (%)	Precision (%)	F1 Score (%)
baseline	63.21	66.84	67.32	67.07	66.81	68.0	69.24	68.61
+satt	68.22	70.31	72.65	71.46	68.81	70.62	78.24	74.23
+gan	75.41	80.73	82.94	81.82	77.33	76.29	85.39	80.58
+satt gan(our)	96.18	95.91	96.32	96.11	94.5	95.68	96.28	95.98

"satt" is the self-attention mechanism, and "gan" is the generative adversarial network.

respectively; Indicators increased by 6 to 7%. In the end, these two key modules were applied in series, not only achieved the highest hit rate of more than 95% on the two data sets, the precision index also exceeded 95 and 96%, and the recall rate was increased to 93.72 and 94.85% of the top level. This fully confirms the important role of attention mechanism and adversarial learning in improving the ability of deep network action recognition, and also highlights the advantages of our improved method in mining multi-modal features. At the same time, we compared and visualized the results in Table 3, as shown in Figure 8.

From the data in Table 4 above, it can be seen that with the continuous optimization of the model structure in our proposed method, the action recognition ability on these two large-scale general-purpose action datasets UCf101 and MultiSports has been greatly improved. Specifically, in comparison with the baseline module, after only adding the self-attention module, the three core evaluation indicators on the MultiSports dataset, namely hit

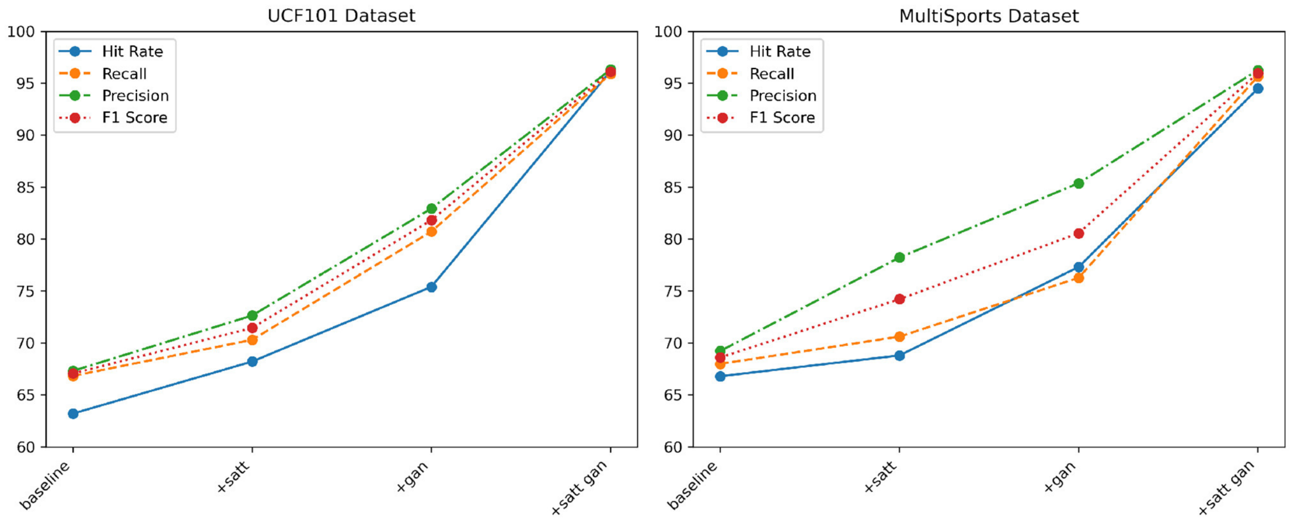


FIGURE 9
Comparison and visualization of Hit Rate, Recall, Precision and F1 Score indicators of different modules based on UCF101 and MultiSports datasets.

rate, recall rate and F1 score, have been improved by more than 2%, respectively; UCf101 dataset The corresponding improvements on the above are even greater, reaching 6.78, 11.76, and 7.15%, respectively, which has verified the role of the attention mechanism in extracting cross-modal correlation features. After adding deep adversarial training on this basis, the improvement of evaluation indicators on the two data sets continues to expand. Among them, the three indicators of the UCf101 dataset all improved within the range of 2 to 9%; the corresponding indicators of the MultiSports dataset increased the most, reaching 10.52, 8.29, and 11.97%, respectively, which further verified how adversarial learning can effectively improve model generalization ability. Finally, the optimization model that integrates attention and confrontation mechanism is adopted, not only makes multiple indicators on UCf101 and MultiSports data sets break through the high level of about 94% for the first time, but also has a recall rate of more than 95.68% on the MultiSports data set; this shows the effectiveness of our method. The optimization effect has achieved generalizability on different types of large-scale action recognition tasks. We compared and visualized the results in Table 4, as shown in Figure 9.

In conclusion, the multimodal deep learning-based robot action recognition method proposed in this study demonstrates significant advantages in experiments conducted on multiple classic volleyball datasets and a large-scale diverse action dataset. By leveraging attention mechanisms to integrate visual and motion features, along with the incorporation of deep adversarial mechanisms to enhance model generalization, the accuracy and recall rate of action recognition have both been notably improved. Particularly, with the integration of the optimized model structure, our method achieves impressive recognition performance across all tested datasets, thus fully validating the reliability and potential of this approach in action recognition tasks.

Through detailed data comparison and analysis, we can clearly witness how the seamless integration of various modules within the model's structure drives the continuous enhancement of recognition capabilities. This not only underscores the correctness of the deep learning architectural approach but also confirms the vital roles of attention mechanisms and adversarial learning in multimodal feature learning. While rooted in the context of volleyball robot requirements, experimental results indicate its promising applicability to other action recognition tasks, further showcasing the method's versatility.

In summary, this work successfully designs and implements a deep multimodal learning algorithm to optimize action recognition capabilities, laying down a methodological foundation for the advancement of robotic sports skills.

4. Conclusion

In preceding chapters, we provided an extensive account of the application of multimodal deep learning methods to enhance robotic cyclic motion skills. In this chapter, we delve into a comprehensive discussion of research outcomes, summarizing key findings from experiments, exploring the significance and contributions of this study, analyzing the strengths and limitations of our approach, and outlining potential avenues for future research.

Through meticulous experimentation and analysis, we observed substantial accomplishments in enhancing robotic skills via multimodal deep learning. The introduction of the cross-modal self-attention mechanism proficiently fuses information from distinct sensors, culminating in comprehensive scene perception. Leveraging Generative Adversarial Networks (GANs) imbues the model with superior data generation and training capabilities, enriching the diversity and practicality of skill training. The implementation of transfer learning further expedites skill augmentation, minimizing the temporal cost of relearning in new environments. The confluence of these modules facilitates remarkable skill enhancement across several pivotal metrics, presenting a positive contribution to the realm of sports robotics.

The significance of this study resides in its insightful and empirical contribution to the progression of cyclic motion robotics. The seamless integration of multimodal perception and deep learning not only elevates robotic prowess in volleyball matches but also ushers in novel prospects for intelligent sports competition and human-robot collaboration. Our research not only theoretically validates this approach but also substantiates its practical efficacy, offering a valuable reference for researchers in related domains.

Throughout this study, we harnessed the inherent advantages of multimodal perception, synergizing information from diverse sensors. This multimodal data processing strategy not only heightens model performance but also enhances robot scene awareness. Simultaneously, our research introduces pivotal technologies such as self-attention mechanisms, GANs, and transfer learning, fully harnessing the potential of deep learning and providing diverse tools and avenues for skill augmentation. However, we acknowledge certain limitations, such as potential model generalization issues stemming from experimental data distribution and the possible challenges and constraints in real-world applications.

Future research directions could encompass the expansion of our approach to diverse sports domains, unraveling the broader potential of multimodal perception and deep learning. Concurrently, optimizing model architectures and algorithms could enhance the efficacy and swiftness of skill augmentation. Furthermore, applying our approach to real volleyball match scenarios could authenticate its viability and efficacy in actual competition. Ultimately, we anticipate our continued research and practical efforts will contribute significantly to the advancement of sports robotics and intelligent sports competition.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

MW: Conceptualization, Data curation, Project administration, Resources, Writing—original draft. ZL: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Writing—original draft, Writing—review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., and Marchand, M. (2014). Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446.* doi: 10.48550/arXiv.1412.4446

Chen, S., Cao, Y., Sarparast, M., Yuan, H., Dong, L., Tan, X., et al. (2020). Soft crawling robots: design, actuation, and locomotion. *Adv. Mater. Technol.* 5, 1900837. doi: 10.1002/admt.201900837

Hong, C., Jeong, I., Vecchietti, L. F., Har, D., and Kim, J.-H. (2021). AI world cup: robot-soccer-based competitions. *IEEE Trans. Games* 13, 330–341. doi: 10.1109/TG.2021.3065410

Hu, Y., Li, Z., and Yen, G. G. (2023). A knee-guided evolutionary computation design for motor performance limitations of a class of robot with strong nonlinear dynamic coupling. *IEEE Trans. Syst. Man Cybernet. Syst.* 53, 4429–4441. doi: 10.1109/TSMC.2023.3249123

Ibrahim, M. S., Muralidharan, S., Deng, Z., Vahdat, A., and Mori, G. (2016). "A hierarchical deep temporal model for group activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1971–1980. Available online at: https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Ibrahim_A_Hierarchical_Deep_CVPR_2016_paper.html

Ji, Y., Li, Z., Sun, Y., Peng, X. B., Levine, S., Berseth, G., et al. (2022). "Hierarchical reinforcement learning for precise soccer shooting skills using a quadrupedal robot," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Kyoto: IEEE), 1479–1486. Available online at: https://ieeexplore.ieee.org/abstract/document/9981984

Kautz, T., Groh, B. H., Hannink, J., Jensen, U., Strubberg, H., and Eskofier, B. M. (2017). Activity recognition in beach volleyball using a deep convolutional neural network: leveraging the potential of deep learning in sports. *Data Mining Knowledge Discov.* 31, 1678–1705. doi: 10.1007/s10618-017-0495-0

Li, B., and Tian, M. (2023). Volleyball movement standardization recognition model based on convolutional neural network. *Comput. Intell. Neurosci.* 2023, 6116144. doi: 10.1155/2023/6116144

Li, R., and Peng, B. (2022). Implementing monocular visual-tactile sensors for robust manipulation. *Cyborg Bionic Syst.* 2022, 9797562. doi: 10.34133/2022/9797562

Li, Y., Chen, L., He, R., Wang, Z., Wu, G., and Wang, L. (2021). "Multisports: a multi-person video dataset of spatio-temporally localized sports actions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Montreal, QC), 13536–13545. Available online at: https://ieeexplore.ieee.org/abstract/document/9711267

Liang, C., and Liang, Z. (2022). The application of deep convolution neural network in volleyball video behavior recognition. *IEEE Access* 10, 125908–125919. doi: 10.1109/ACCESS.2022.3221530

Mi, Z., Jiang, X., Sun, T., and Xu, K. (2020). Gan-generated image detection with self-attention mechanism against gan generator defect. *IEEE J. Select. Top. Signal Process.* 14, 969–981. doi: 10.1109/JSTSP.2020.2994523

Niu, Z., Zhong, G., and Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing* 452, 48–62. doi: 10.1016/j.neucom.2021.03.091

Olaniyan, O. T., Adetunji, C. O., Adeniyi, M. J., and Hefft, D. I. (2022). "Computational intelligence in iot healthcare," in *Deep Learning, Machine Learning and IoT in Biomedical and Health Informatics* (CRC Press), 297–310. Available online at: https://www.taylorfrancis.com/chapters/edit/10.1201/9780367548445-19/computational-intelligence-iot-healthcare-olugbemi-olaniyan-charles-adetunji-mayowa-adeniyi-daniel-ingo-hefft

Oliff, H., Liu, Y., Kumar, M., Williams, M., and Ryan, M. (2020). Reinforcement learning for facilitating human-robot-interaction in manufacturing. *J. Manufact. Syst.* 56, 326–340. doi: 10.1016/j.jmsy.2020.06.018

Oliveira, L. S., Moura, T. B. M. A., Rodacki, A. L. F., Tilp, M., and Okazaki, V. H. A. (2020). A systematic review of volleyball spike kinematics: implications for practice and research. *Int. J. Sports Sci. Coach.* 15, 239–255. doi: 10.1177/1747954119899881

Salim, F. A., Haider, F., Tasdemir, S. B. Y., Naghashi, V., Tengiz, I., Cengiz, K., et al. (2019). "Volleyball action modelling for behavior analysis and interactive multi-modal feedback," in *Proceedings of the 15th International Summer Workshop on Multimodal Interfaces* (Ankara), 50.

Siedentop, D., and Van der Mars, H. (2022). *Introduction to Physical Education, Fitness, and Sport*. Human Kinetics.

Siegel, J., and Morris, D. (2020). "Robotics, automation, and the future of sports," in *21st Century Sports: How Technologies Will Change Sports in the Digital Age*, ed S. L. Schmidt (Springer), 53–72.

So, J., Kim, U., Kim, Y. B., Seok, D.-Y., Yang, S. Y., Kim, K., et al. (2021). Shape estimation of soft manipulator using stretchable sensor. *Cyborg Bionic Syst.* 2021, 9843894. doi: 10.34133/2021/9843894

Soomro, K., Zamir, A. R., and Shah, M. (2012). UCF101: a dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402.*

Tang, J. (2021). An action recognition method for volleyball players using deep learning. *Sci. Prog.* 2021, 1–9. doi: 10.1155/2021/3934443

Thuruthel, T. G., Shih, B., Laschi, C., and Tolley, M. T. (2019). Soft robot perception using embedded soft sensors and recurrent neural networks. *Sci. Robot.* 4, eaav1488. doi: 10.1126/scirobotics.aav1488

Wang, H., Sahoo, D., Liu, C., Shu, K., Achananuparp, P., Lim, E.-P., et al. (2021). Cross-modal food retrieval: learning a joint embedding of food images and recipes with semantic consistency and attention mechanism. *IEEE Trans. Multimedia* 24, 2515–2525. doi: 10.1109/TMM.2021.3083109

Weiss, A., Wortmeier, A.-K., and Kubicek, B. (2021). Cobots in industry 4.0: a roadmap for future practice studies on human–robot collaboration. *IEEE Trans. Hum. Mach. Syst.* 51, 335–345. doi: 10.1109/THMS.2021.3092684

Wenninger, S., Link, D., and Lames, M. (2020). Performance of machine learning models in application to beach volleyball data. *Int. J. Comput. Sci. Sport* 19, 24–36. doi: 10.2478/ijcss-2020-0002

Xia, H., Tracy, R., Zhao, Y., Fraisse, E., Wang, Y.-F., and Petzold, L. (2022). "VREN: volleyball rally dataset with expression notation language," in *2022 IEEE International Conference on Knowledge Graph (ICKG)* (Orlando, FL: IEEE), 337–346.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2019). "Modeling tabular data using conditional gan," in *Advances in Neural Information Processing Systems 32.* Available online at: https://proceedings.neurips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html

Zhang, J. (2019). Gradient descent based optimization algorithms for deep learning models training. *arXiv preprint arXiv:1903.03614.* doi: 10.48550/arXiv.1903.03614

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., et al. (2020). A comprehensive survey on transfer learning. *Proc. IEEE* 109, 43–76. doi: 10.1109/JPROC.2020.3004555

frontiers | Frontiers in Neurorobotics

TYPE Original Research
PUBLISHED 14 November 2023
DOI 10.3389/fnbot.2023.1280501

OPEN ACCESS

EDITED BY
Long Jin,
Lanzhou University, China

REVIEWED BY
Lin Chen,
Chinese Academy of Sciences (CAS), China
Alexander Rassau,
Edith Cowan University, Australia

*CORRESPONDENCE
Wenshuai Wang
✉ wangws@stu.pku.edu.cn

RECEIVED 20 August 2023
ACCEPTED 30 October 2023
PUBLISHED 14 November 2023

CITATION
Wang H, Wang W, Li W and Liu H (2023) Dense captioning and multidimensional evaluations for indoor robotic scenes.
Front. Neurorobot. 17:1280501.
doi: 10.3389/fnbot.2023.1280501

Dense captioning and multidimensional evaluations for indoor robotic scenes

Hua Wang[1,2], Wenshuai Wang[1]*, Wenhao Li[1] and Hong Liu[1]

[1]Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, Shenzhen, China,
[2]School of Artificial Intelligence, Hebei University of Technology, Tianjin, China

The field of human-computer interaction is expanding, especially within the domain of intelligent technologies. Scene understanding, which entails the generation of advanced semantic descriptions from scene content, is crucial for effective interaction. Despite its importance, it remains a significant challenge. This study introduces RGBD2Cap, an innovative method that uses RGBD images for scene semantic description. We utilize a multimodal fusion module to integrate RGB and Depth information for extracting multi-level features. And the method also incorporates target detection and region proposal network and a top-down attention LSTM network to generate semantic descriptions. The experimental data are derived from the ScanRefer indoor scene dataset, with RGB and depth images rendered from ScanNet's 3D scene serving as the model's input. The method outperforms the DenseCap network in several metrics, including BLEU, CIDEr, and METEOR. Ablation studies have confirmed the essential role of the RGBD fusion module in the method's success. Furthermore, the practical applicability of our method was verified within the AI2-THOR embodied intelligence experimental environment, showcasing its reliability.

KEYWORDS

indoor robotic scene, dense captioning, RGBD fusion, multidimensional evaluation, top-down attention

1 Introduction

As artificial intelligence technology continues to evolve, mobile robots are taking on increasingly pivotal roles across a multitude of fields (Rubio et al., 2019; Huang et al., 2020; Liu et al., 2022). To enable these robots to more effectively comprehend and adapt to complex, ever-changing indoor environments, it becomes essential to provide a detailed description of the scene (Johnson et al., 2016; Chen et al., 2021). This involves extracting semantic information—such as objects, attributes, and relationships within the scene—and articulating it in natural language. By doing so, we can significantly enhance a robot's perceptual and interactive capabilities, thereby elevating its level of intelligence and the overall user experience (Sheridan, 2016). The task of providing semantic descriptions of scenes is of paramount importance, as it is key to facilitating effective interaction between robots and humans, and crucial to a robot's understanding of human needs.

Scene description refers to the ability of machines to generate high-level natural language descriptions based on given scene images. Several Scene Description methods have been developed for indoor scenes, with a recent focus on Dense Captioning based on 3D point clouds. In 2021, Chen et al. (2021) proposed an end-to-end method called Scan2Cap, which effectively locates and describes 3D objects in the ScanRefer dataset and extracts spatial relationships within the scene. Yuan et al. (2022) introduced a cross-modal Transformer model, X-Trans2Cap, which integrates features from auxiliary 2D modalities into point clouds through knowledge distillation, achieving great performance improvement in this task. Jiao et al. (2022) proposed a multi-level relationship mining model called MORE, aiming to improve 3D Dense Captioning by capturing and utilizing complex relationships within 3D scenes.

The task of providing dense scene captioning presents numerous challenges (Cai et al., 2022). To begin with, in the context of 2D scene captioning, the input from a single modality is often insufficient, making it difficult to discern when objects are occluded or when the viewpoint within the scene changes. Additionally, while 3D scene captioning can capture comprehensive scene information, the computational cost of performing convolution and attention operations on point cloud data is high, and there is an abundance of sparse, irrelevant information. Ultimately, the existing methods of RGBD input have not effectively utilized the information available in depth images, which serves as the motivation for this research. We want to implement a method that could reduce the amount of computation while expressing spatial relationships better, so we came up with RGBD2Cap.

The main contribution of this paper includes the following three aspects: Firstly, we propose a feature extraction method based on RGB+D image multimodal fusion. This method, which is grounded in the transformation between 3D point clouds and 2D images, is combined with a semantic captioning generation module to form RGBD2Cap. Secondly, we design and implement a multi-dimensional evaluation method for scene semantic captioning. This includes both manual and automatic evaluations, and utilizes simulation scenes to assess the model within an embodied intelligence experimental environment. Lastly, the model presented in this article has achieved the highest accuracy according to our evaluation metrics.

2 Related work

2.1 2D image and scene captioning

Since its introduction by Johnson et al. (2016), dense captioning has emerged as a subfield of image captioning, with the encoder-decoder architecture becoming the prevailing solution (Cho et al., 2014).

Initial approaches (Mao et al., 2014) to dense image captioning using the encoder-decoder architecture combined Convolutional Neural Networks (CNNs) (LeCun et al., 2015) and Long Short-Term Memory (LSTM) networks (Xu et al., 2015). These methods used the image feature vector extracted by the CNN as the LSTM's initial state and generated descriptive statements word by word.

With the rise of attention mechanisms in natural language processing, methods (Xu et al., 2015; Anderson et al., 2018) combining CNNs and attention mechanisms have emerged. These methods dynamically select the most relevant region feature vectors at each time step based on the current generation state, combining them with global feature vectors as input to subsequent language generation models such as LSTM or Transformer.

Yang et al. (2017) introduced a method that combines joint inference and contextual information fusion to address two significant challenges in the current image-intensive description task. This approach generates improved descriptions by emphasizing visual cues from surrounding salient image regions as contextual features. Kim et al. (2019) introduced a new task, "Relation Captioning," which generates multiple captions for relational information between objects in an image. They utilized a multi-task triple stream network (MTTSNet) that captures the relational information between detected objects, providing precise concepts and rich representations.

2.2 3D scene captioning

3D vision has become increasingly popular in recent years (Qi et al., 2017; Li et al., 2022; Shao et al., 2022), and 3D detection methods performed on point clouds are becoming more common in 3D vision research.

Chen et al. (2021) pioneered the task of dense captioning in RGB-D scans, a field that has yet to fully explore the discriminative description of objects in complex 3D environments. Yuan et al. (2022) furthered this research by investigating a cross-modal knowledge transfer using a Transformer for 3D dense captioning. Their model, X-Trans2Cap, leverages a teacher-student framework for knowledge distillation to enhance the performance of single-modal 3D captioning.

In the spirit of neural machine translation, Wang et al. (2022) proposed SpaCap3D. This model features a spatiality-guided encoder and an object-centric decoder, both of which contribute to the generation of precise and spatially-enhanced object captions.

However, existing methods often overlooking contextual information such as non-object details and background environments within point clouds. To address this, Zhong et al. (2022) utilized point cloud clustering features as contextual information, incorporating non-object details and background environments into the 3D dense captioning task.

Jiao et al. (2022) aimed to improve 3D dense captioning by capturing and utilizing complex relations within the 3D scene. They proposed MORE, a Multi-Order RElation mining model, to generate more descriptive and comprehensive captions. Chen et al. (2022) introduced UniT3D, a fully unified transformer-based architecture for jointly solving 3D visual grounding and dense captioning.

Although the representation of 3D point cloud scenes has achieved considerable performance to some extent, its computational overhead remains excessively large. This is primarily due to the sparsity of the 3D point cloud information, which impedes the efficient utilization of features. This paper proposes a method based on RGBD static images, effectively integrating

RGB and Depth features. While reducing computational load, this approach also ensures the model's acquisition of spatial information, thereby enhancing the accuracy of the generated descriptions.

3 Proposed method

The research of this paper is to train a deep learning model based on the RGBD images corresponding to indoor 3D scenes, so that it can automatically generate the corresponding linguistic descriptions. In order to accomplish these goals, this paper accomplish the following specific tasks. First, we need to pre-process the original point cloud data to obtain 2D and depth images corresponding to different objects in the scene. Then, we design a RGB and Depth multimodal feature extraction network to extract and fuse the features of RGB and depth images. In addition, we need a target detection network to detect the objects in the scene images so that the subsequent Top-down Attention LSTM model can accurately understand the objects in the images. Finally, the features extracted by the neural network are fed into the text generation network to generate text for the purpose of understanding the high-level semantic information of the scene. The overall structure of the proposed method is shown in Figure 1.

3.1 Rendering of 3D scenes

This study employs the ScanRefer (Chen et al., 2020) dataset for model training, which is an extension of the ScanNet dataset with added high-level semantic descriptions. ScanNet provides a rich array of indoor 3D scene meshes, semantic labels, and 2D video frame images with corresponding depth maps. However, we refrain from using ScanNet's 2D image data directly for training due to the blurriness of most images, which hampers effective capture of the scene's visual information. Instead, we use the viewpoints provided by the ScanRefer dataset to render the 3D data, yielding clearer 2D data.

The rendering process of the 3D scene adheres to the principle of camera projection (Kannala and Brandt, 2006). It begins with transforming the scene points in the world coordinate system using the camera's external parameter matrix, yielding their coordinates in the camera's coordinate system. These points are then converted to the image coordinate system using the camera's internal parameter matrix.

The initial step involves the transformation from the world coordinate system to the camera coordinate system, a rigid transformation composed of translation and rotation. In this study, a right-hand coordinate system is used for world coordinates. If a point in the scene has coordinates (x, y, z) in the world coordinate system. We aim to obtain its coordinates (x', y', z') in the camera coordinate system, this can be achieved through the following matrix transformation:

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \left[\begin{array}{ccc|c} R_{11} & R_{12} & R_{13} & t_x \\ R_{21} & R_{22} & R_{23} & t_y \\ R_{31} & R_{32} & R_{33} & t_z \end{array}\right] \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}, \tag{1}$$

where t represents the translation vector of the point coordinates, and the orthogonal matrix R represents the rotation matrix of the point's coordinates in space. The values of both are determined by the position of the camera in the world coordinate system and the direction of the optical axis. The external parameter matrix of the camera is composed of the rotation matrix R and the translation vector t, represented as $[R|t] \in R^{3\times4}$.

Next is the transformation from the camera coordinate system to the normalized device coordinate system, which is usually achieved through perspective projection. For a point (x', y', z') in the camera coordinate system, the following matrix transformation can be used to describe this process:

$$\begin{bmatrix} x'' \\ y'' \\ z'' \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix}, \tag{2}$$

where f represents the focal length of the camera, and (x'', y'', z'') are the coordinates of the point in the normalized device coordinate system. This transformation maps the 3D points in the camera coordinate system to a 2D, while preserving the depth information of each point.

Finally, there is the transformation from the normalized device coordinate system to the image coordinate system, which can be achieved through the simple scaling and offset. For a point (x'', y'', z'') in the normalized device coordinate system, we want to obtain its coordinates (u, v) in the image coordinate system, which can be achieved through the following formula:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} w/2 & 0 \\ 0 & h/2 \end{bmatrix} \begin{bmatrix} x''y'' \end{bmatrix} + \begin{bmatrix} w/2 \\ h/2 \end{bmatrix}, \tag{3}$$

where w and h represent the width and height of the image, respectively. This transformation maps the points in the normalized device coordinate system to the image coordinate system, generating the final 2D image.

The above is the whole process we used to convert the point cloud in the scene, from the world coordinate system to the image coordinate system. The whole process is linear and can be achieved by a series of matrix multiplications. This allows us to obtain a mapping of the 3D point cloud data onto the 2D image, which can then be processed and analyzed using 2D image processing techniques.

3.2 RGB and depth multimodal fusion networks

The network accepts an RGB image and a depth image as inputs. Its architecture is grounded in ResNet101 (He et al., 2015), a deep residual network of 101 convolutional neural network layers. This network addresses the issues of vanishing and exploding gradients, common in deep neural network training, through residual learning.

The feature fusion approach employed in this network is a third-branch multilevel fusion, as shown in Figure 2. Specifically, we start with the feature map generated by the third convolutional

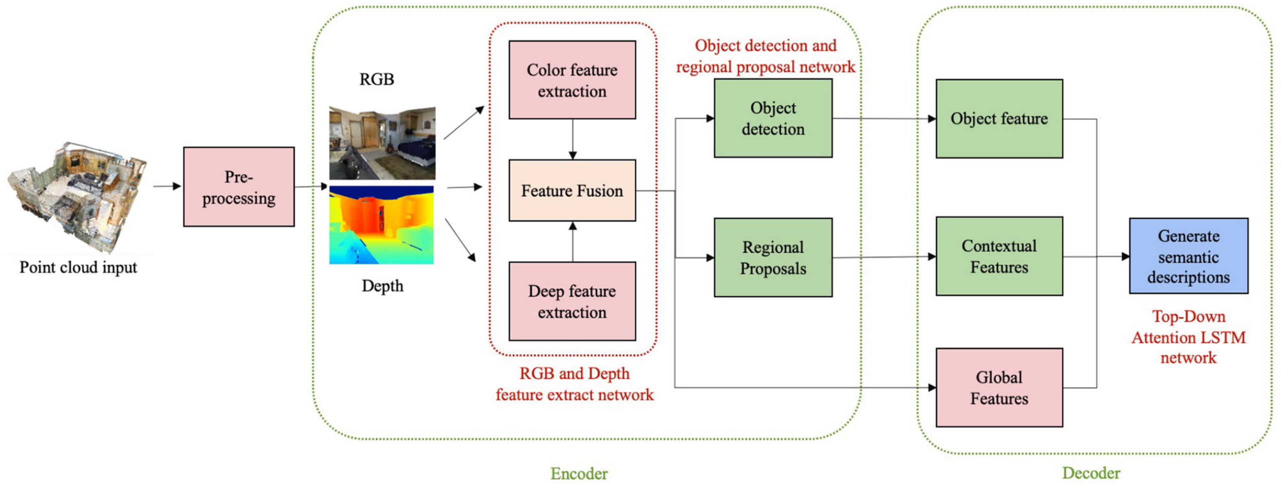


FIGURE 1
The general structure of the proposed method.

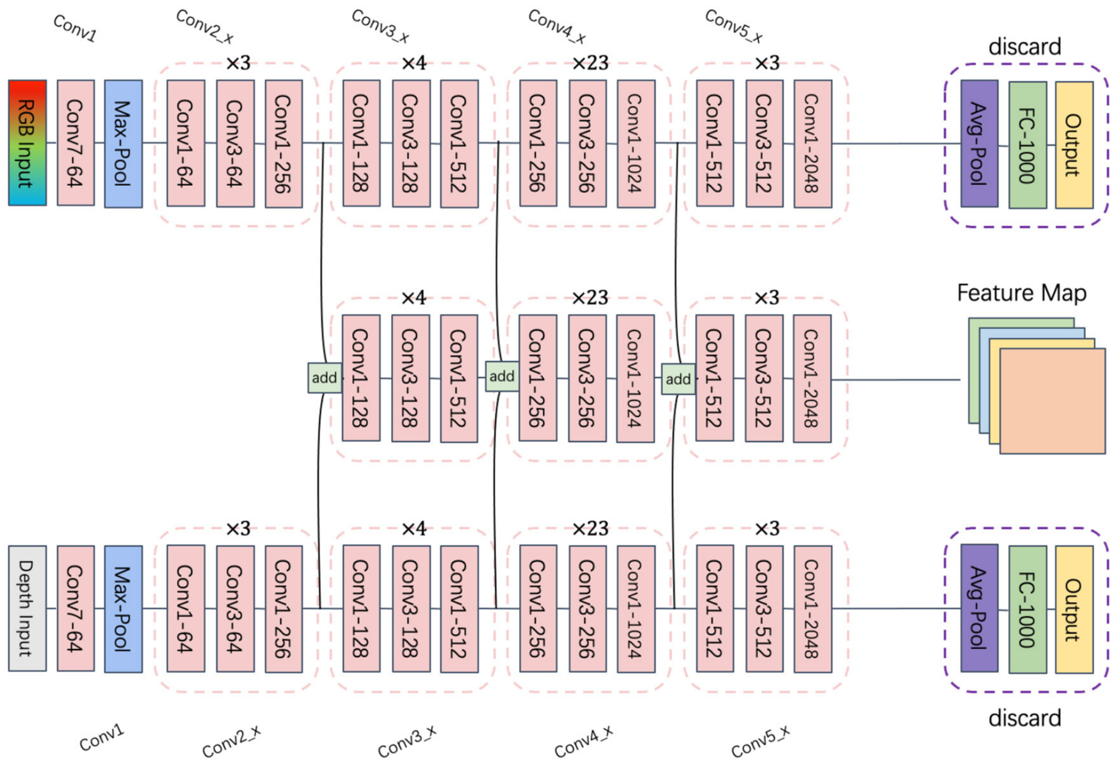


FIGURE 2
RGB and Depth multimodal fusion networks.

layer of ResNet101. The RGB and depth feature maps from this convolutional layer are summed and fused separately to form the network's third branch. The same convolutional operation is performed on this third branch, and the feature maps obtained from subsequent convolutional layers are continuously added to yield the final RGBD multimodal features.

Our feature extraction network is bifurcated into two branches: the RGB branch and the depth branch. The RGB image and the depth image are processed through their respective convolution layers to extract features and generate their individual feature maps. These two feature maps are then fused using the feature fusion method to obtain RGBD multimodal features, which serve as the third branch for multilevel fusion. This network omits the final fully-connected and softmax layers of ResNet, bypassing classification result output and directly utilizing its feature maps for subsequent tasks.

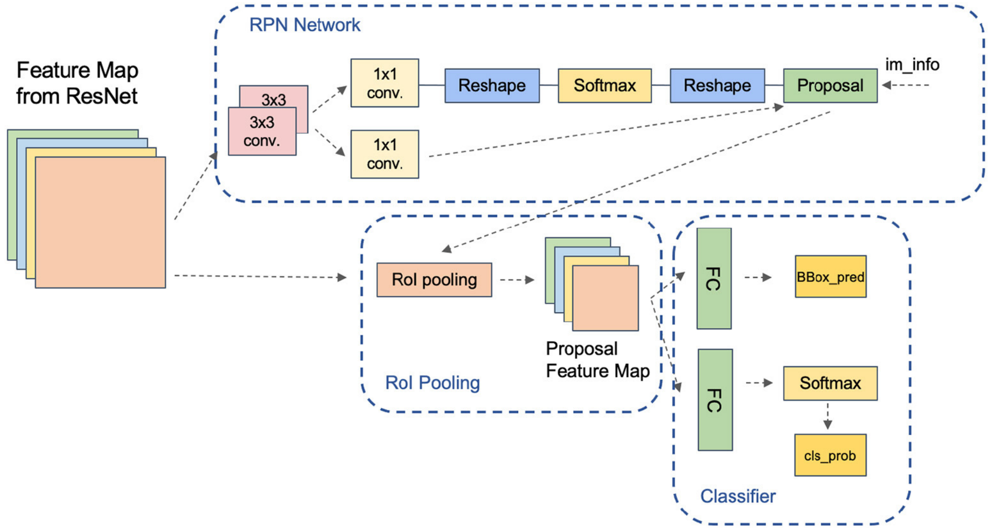


FIGURE 3
Target detection and region proposal network.

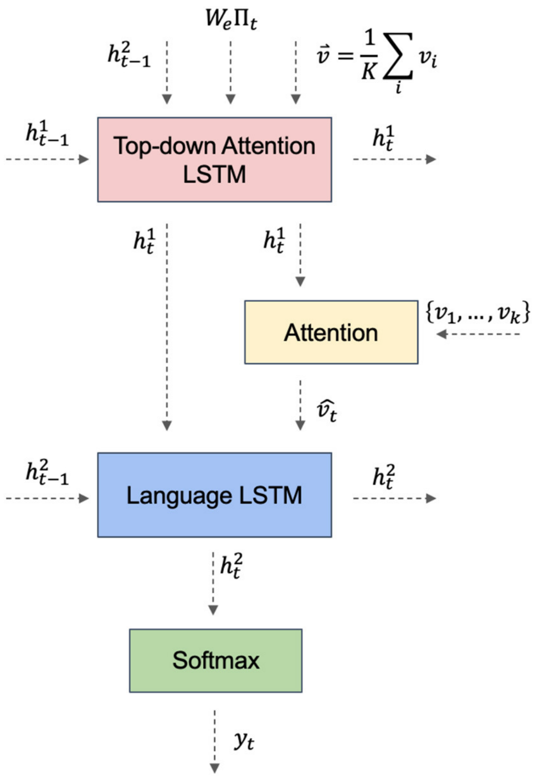


FIGURE 4
Top-down attention LSTM network.

3.3 Target detection and region proposal network

The Bottom-Up and Top-Down Attention model (Anderson et al., 2018) comprises two components: a bottom-up image feature extractor and a top-down language generator. The bottom-up image feature extractor employs a Faster-RCNN (Ren et al., 2015) detector to identify a set of potential visual regions, generating a fixed-length feature vector for each region.

As shown in Figure 3, this study employs a Faster-RCNN-based object detection and region proposal network, utilizing the previously mentioned multimodal fusion ResNet101 as its backbone, augmented with an RPN network and an RoI Pooling layer. The RPN network, which is fully convolutional, generates candidate bounding boxes. It takes the output feature map of the backbone network as input and produces a series of candidate bounding boxes along with their corresponding scores. A 3 × 3 convolution generates scores for each position, and non-maximum suppression is applied to eliminate overlapping candidate boxes. The RoI Pooling layer takes the output feature map of the backbone network and a series of candidate boxes as input, outputting a fixed-size feature vector after pooling. The final pooling results are concatenated to form the ultimate feature vector.

3.4 Top-down attention LSTM network

The top-down language generator in the Bottom-Up and Top-Down Attention model employs an attention mechanism as shown in Figure 4. This mechanism uses the currently generated word as a query, calculates its similarity with the bottom-up feature vector, and produces a set of attention weights. These weights are then used to compute a weighted average of each feature vector, which is used to generate the next word.

The top-down attention mechanism is the heart of the model. The model uses the currently generated word as a query at each time step, calculates its similarity with the bottom-up feature vector, and produces a set of attention weights. These weights are then used to compute a weighted average of each feature vector, which is used to generate the next word. This attention mechanism can be viewed

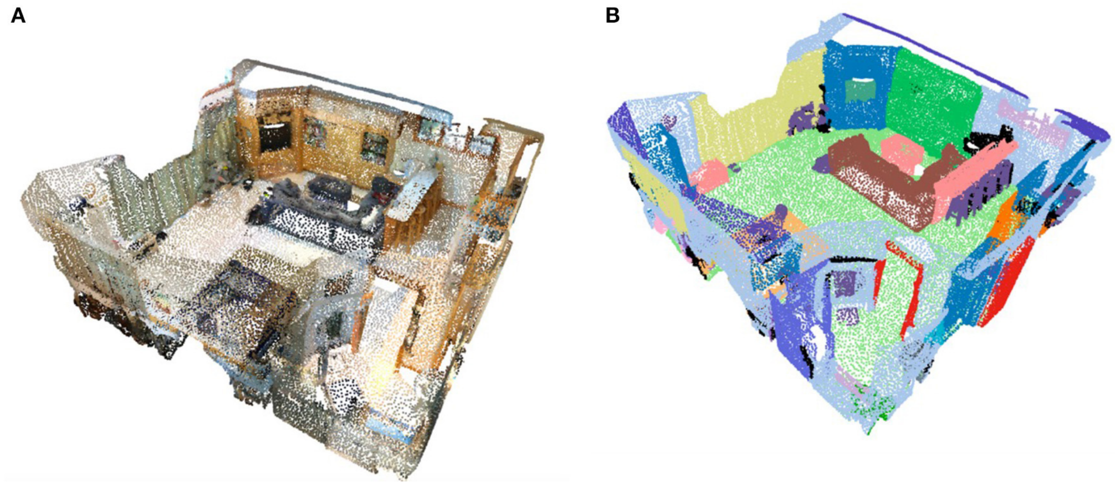


FIGURE 5
(A) Visualization of point cloud data. **(B)** Visualization of the labels of point cloud data.

as a top-down interpretation of the image, integrating the generated language with the underlying image representation to produce a more precise image description.

In this module, the global features, object features, and context features obtained from the previous networks are fused, and to utilize these three features effectively, we use the following method for fusion. Firstly, the global and target features with the same dimension are spliced and fused, then a fully connected network with an activation function is used to scale the fused features to the same dimension as the contextual features, and then they are spliced twice to get the final fused features, which can effectively utilize the extracted contextual features.

4 Experiments

4.1 Dataset

Generating scene descriptions for robots necessitates a computer vision approach that can convert environmental data into natural language descriptions. Several datasets have been developed to provide high-level language descriptions for various scenes, including the ScanRefer dataset.

ScanRefer (Chen et al., 2020) is a dataset designed explicitly for dense scene descriptions, primarily used in robotic indoor scene understanding tasks. It provides semantic scene description information, facilitating robots' comprehension of their surroundings. The dataset comprises 800 annotated scenes, 11,046 stereo location frames of objects, and 51,583 corresponding textual descriptions. It offers not only a wealth of scene description data but also high-quality 3D scene data. By employing 3D projection, we can map the objects in the scene onto a 2D plane, making it suitable for the RGBD2Cap model presented.

ScanRefer builds upon the ScanNet (Dai et al., 2017) dataset by adding natural language descriptions. As shown in Figures 5A, B, ScanNet provides 3D point clouds and their corresponding semantic labels, resulting from high-quality scene reconstruction. In this study, we utilize the 3D data from the dataset and select viewpoints provided by ScanRefer to render the point cloud scenes. The authors of ScanRefer provide viewpoint information for different camera locations in each scene in the Annotated viewpoints file. This information includes the camera location, rotation angle, and look at (the point the camera is currently aimed at), which we use to set the camera pose.

4.2 Rendering of 2D images

The rendering of the 3D scene using Pytorch3D (Ravi et al., 2020) is shown in Figure 6. From left to right, the RGB color image of a viewpoint, the rendered image with labels, and the depth image are shown.

4.3 Configuration of the training model

This study utilized the Python programming language and the PyTorch deep learning framework to implement the algorithm. The hardware setup for the experiment included a NIVIDA Tesla P100 GPU (16GB), 80GB of RAM, and 70GB of available disk space. The software environment was configured with Ubuntu 18.04, Python 3.8, Cuda 11.1, and PyTorch 1.8.1.

The experimental procedure began with the fusion of the ScanRefer dataset with RGBD images to extract image features. The primary architecture used in the training process was a convolutional neural network and a long short-term memory network. The model was trained using the Adam optimizer, with a batch size of 14 and 100 epochs. The initial learning rate was set at 0.0005, and a weight decay parameter of 0.0001 was used to control model complexity. Intersection over Union (IOU) thresholds were set at 0, 0.25, and 0.5. The number of sampled point clouds was 40,000, with 562 scenes in the training set and 141 in the validation

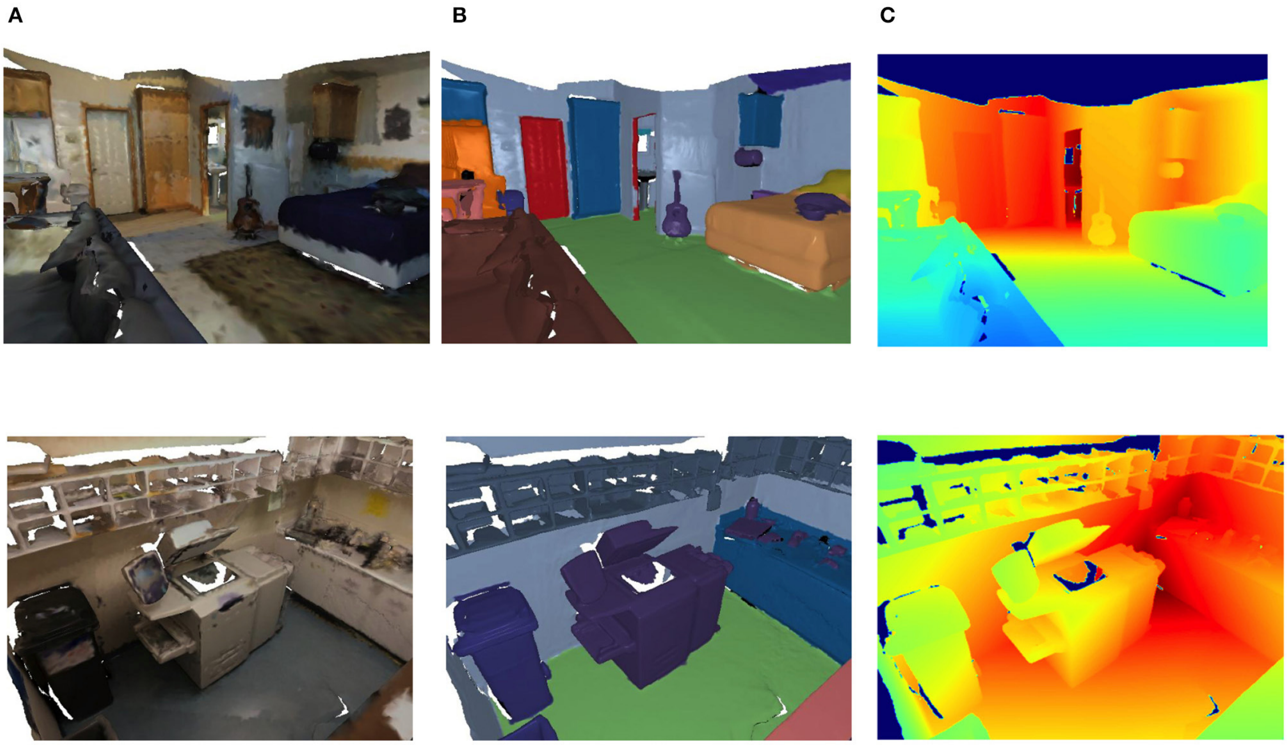


FIGURE 6
Multi-view image based on pytorch3d rendering. **(A)** RGB image. **(B)** Labeled image. **(C)** Depth image.

set. After rendering, the training set comprised 36,665 samples, and the validation set included 9,508 samples.

The loss of the RGBD2Cap network is a multi-task loss, including target detection loss and semantic description loss. The loss for target detection includes classification loss and bounding box regression loss, while the text generation part can directly use the cross-entropy loss of text prediction probability. The final multi-task loss value at the end of model training was 0.26.

4.4 Scene dense captioning and evaluation methods

4.4.1 Metrics-based evaluation

The objective of the dense captioning task is to identify and articulate all objects and events of interest within an image. This task merges two subtasks: object detection and image captioning. Consequently, its evaluation metrics are a fusion of the metrics used for these two subtasks.

Firstly, the Mean Average Precision (mAP) is typically used as the evaluation metric for object detection. The mAP represents the Area Under Curve (AUC) of the average precision-recall curve across all categories. For each category, detections are ranked based on their predicted confidence, followed by the calculation of precision and recall. The precision-recall curve is then plotted, and the area under it is calculated to obtain that category's Average Precision (AP). The final mAP is obtained by averaging the AP across all categories.

Secondly, image captioning is evaluated using metrics such as BLEU, CIDEr, Meteor, and Rouge. BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) assesses the similarity between generated and reference descriptions primarily through *n*-gram accuracy. CIDEr (Consensus-based Image Description Evaluation) (Vedantam et al., 2015) gauges the quality of descriptions by calculating the TF-IDF-weighted cosine similarity between generated descriptions and a set of reference descriptions. Meteor (Metric for Evaluation of Translation with Explicit ORdering) (Banerjee and Lavie, 2005) and Rouge (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004) evaluate description quality by computing the longest common subsequence between generated and reference descriptions.

In dense captioning tasks, these evaluation metrics for object detection and image captioning are typically used in conjunction. Specifically, mAP is used to assess the model's performance on the object detection task, while BLEU, CIDEr, Meteor, and Rouge are used to evaluate the model's performance on the image captioning task. Finally, these evaluation metrics can be combined in a weighted manner to derive a comprehensive evaluation metric for assessing the model's overall performance on the dense captioning task.

In this paper, we evaluate the completed training model and obtain several evaluation metrics data, including (BLEU1-4, cider, mAP@0.5, meteor, rouge, and many other evaluation metrics). The

TABLE 1 Algorithm comparison and ablation study on RGBD2Cap components.

	BLEU-4	CIDEr	ROUGE-L	METEOR
RGB (DenseCap) (Johnson et al., 2016)	20.1	32.7	38.2	21.0
RGB (without fusion)	20.7	34.5	41.6	22.9
Show and tell (without attention)	18.3	33.5	**46.9**	21.09
RGB+D fusion (ours)	**21.5**	**35.1**	38.8	**23.3**

Bold values indicate the optimal value of each method for a given evaluate metric.

TABLE 2 Comparison of time and accuracy between 2D and 3D methods.

	BLEU-4	CIDEr	ROUGE-L	METEOR	Train time (h)
RGBD2Cap (ours)	21.5	35.1	38.8	**23.3**	**8**
Scan2Cap (Chen et al., 2021)	**23.32**	**39.08**	**44.78**	21.97	71

Bold values indicate the optimal value of each method for a given evaluate metric.

IOU thresholds *k* in the data table are all taken as 0.5. The results are shown in Table 1.

Since no experimental studies are based on RGBD fusion so far, the proposed model is compared with the algorithm without RGBD fusion.

The RGB(DenseCap) row in Table 1 uses the rendered RGB image as input, and the Dense Captioning of the scene is obtained by using the method in paper[]. The RGB(Without Fusion) line also takes the same image as input and uses the RGBD2Cap network without the Depth branch and the third branches to get the DenseCap. The last row in Table 1 is our complete proposed RGBD2Cap method. Based on the data in the table, it can be seen that the performance of the proposed model is optimal in the three indexes of BLEU-4, CIDEr, and METEOR, which can verify the effectiveness of the RGBD fusion module.

Furthermore, ablation experiments were conducted to ascertain the effectiveness of the Top-down Attention and FasterRCNN modules. As depicted in Table 1, the model's performance across all three metrics declines when the Attention module is not utilized, indicating the module's crucial role in feature extraction during semantic description generation.

In addition, we compare the proposed method RGBD2Cap with the 3D method Scan2Cap (Chen et al., 2021), and the obtained results are shown in Table 2. Both methods are trained on the ScanRefer dataset, the difference is that RGBD2Cap uses a rendered RGBD image as the input to the model, while Scan2Cap directly uses a 3D point cloud as the input. Both models are trained on a 2080Ti GPU for 50 epochs to ensure fairness. Based on the experimental results, it can be learned that although the 3D model outperforms our method in the three metrics, its training time is 9 times longer than that of RGBD2Cap, greatly shortening the training time while reducing the performance loss.

TABLE 3 Performance of using Faster-RCNN as a target detector vs. real bounding box to generate description results.

	BLEU-4	CIDEr	ROUGE-L	METEOR
Faster-RCNN	21.5	35.1	38.8	23.3
Ground truth	**24.3**	**35.7**	**39.3**	**23.5**

Bold values indicate the optimal value of each method for a given evaluate metric.

Lastly, we verify the impact of the Faster-RCNN module's detection capabilities on the description performance by contrasting it with the actual bounding box, as shown in Table 3. The features extracted using the real bounding box of the object are more precise, hence the semantic description based on it will also yield more accurate descriptions. Following experimental verification, it was found that the model exhibits a slight decrease in the four indicators. Still, the decrease is minimal, thus affirming the feasibility of the end-to-end model. The target features produced using Faster-RCNN as the target detector and feature box extractor serve as a solid foundation for semantic description.

4.4.2 Manual evaluation

Because the high-level semantics are more difficult to describe formalistically, manual evaluation is essential, and this paper next evaluates a manual sample of training results.

A randomly selected sample from the validation set was used for inference prediction, and the results are presented in Figure 7. The captioning of the red box is *"The chair is brown. It is to the left of the desk"*, in which the object's color information and spatial location are accurately displayed; the captioning of the white box is *"The monitor is on the desk on the right side. It is the monitor that is closest to the window"*, although the real label of the computer on the desktop is "laptop", the object vocabulary "monitor" given in the description is similar; this description shows very detailed spatial location information; the captioning of the green box is *"The desk is on the right side of the room. There is a chair in front of the desk."* This description shows the position of the desk object in the room and accurately expresses its spatial relationship with the chair in front of it.

However, not all scenes are accurately described, and Figure 8 shows another randomly selected sample from the validation set. The captioning of the red box in the figure is *"This is a white pillow. It is on a gray couch."* Although the object's color is accurately described as white, the white bed sheet is mistakenly identified as a pillow and the bed below as a sofa, which is a misjudgment. The text of the blue box is *"This is a brown nightstand. It is next to a bed"*, which accurately shows that the object is a brown nightstand; it also points out that its orientation is next to the bed; the text of the pink box is *"this is a radiator. It sets along the wall."* This sentence incorrectly identifies the object as a radiator, probably because the picture shows an incomplete object, but it correctly conveys that the object is against the wall.

From the results, it can be seen that the current field still faces many challenges, and future research directions could be more fine-grained feature extraction to achieve a more accurate description.

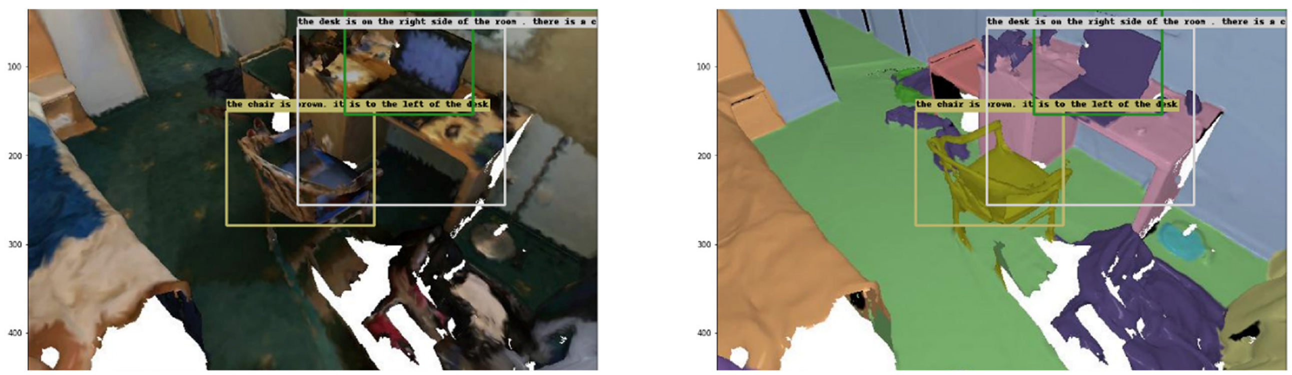

the desk is on the right side of the room . there is a c
the chair is brown. it is to the left of the desk
100
200
300
400

FIGURE 7
Example 1 of dense captioning results in the validation set.

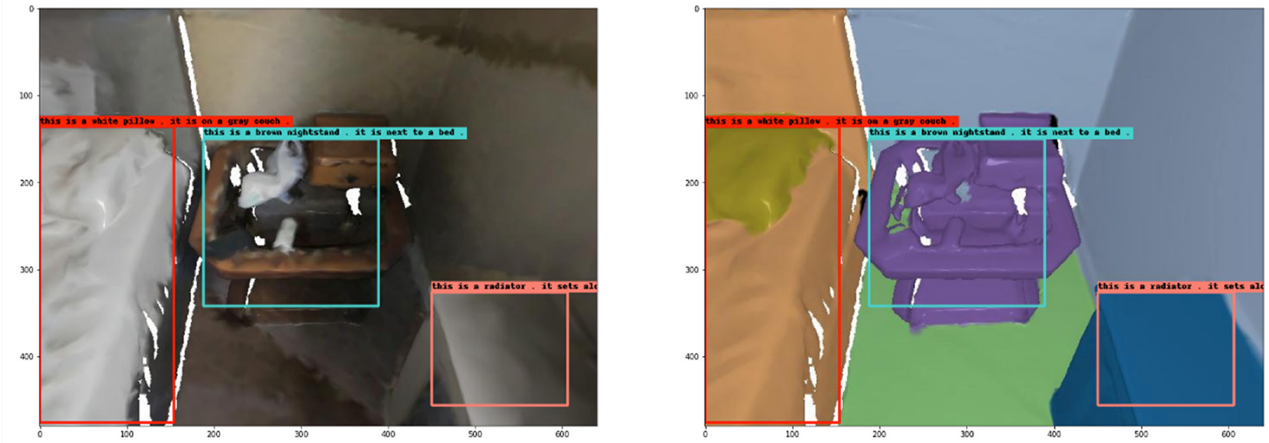

this is a white pillow . it is on a gray couch .
this is a brown nightstand . it is next to a bed .
this is a radiator . it sets alc
0
100
200
300
400
500
600

FIGURE 8
Example 2 of dense captioning results in the validation set.

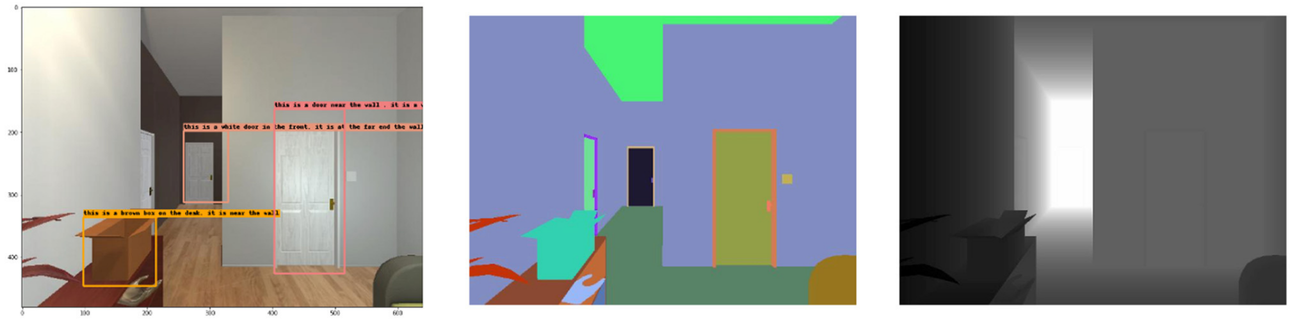

this is a door near the wall . it is a w
this is a white door in the front. it is at the far end the wall
this is a brown box on the desk. it is near the wall

FIGURE 9
RGB, Labeled, and Depth images of scenes in AI2-THOR environment.

4.5 Simulation tests in AI2-THOR

4.5.1 AI2-THOR

AI2-THOR is an embodied AI experimental environment designed to simulate real-world environments to train and test AI systems (Kolve et al., 2017; Deitke et al., 2020). This simulation environment contains a variety of detailed indoor scenarios such as kitchen, bedroom, bathroom, and living room. In AI2-THOR, AI intelligence can explore and interact with the environment through a series of actions, such as moving, viewing, grasping, and manipulating objects. This design allows the intelligent body to learn and understand the properties and relationships of objects in the environment and how they affect the execution of tasks as it performs them.

A key feature of AI2-THOR is its support for scene semantics, for which objects are provided with labels with semantic information. In this paper, RGBD2Cap is further evaluated by controlling the actions of the intelligence in AI2-THOR, acquiring single frames of images in the scene and their depth images as input samples for the model, and observing the correlation between the model's output and the images.

4.5.2 Operation details

The operation of AI2-THOR is facilitated through Python, with the research team providing a Python API for public experimentation. Initially, the AI2-THOR experimental environment is installed and initialized, typically involving the selection of a scene (e.g., kitchen, bedroom, etc.) and establishing the AI agent's initial position and orientation. Once the environment is initialized, the agent is primed to commence action execution.

The system's "move" and "rotate" actions can be utilized to capture a single frame from varying scene perspectives. For instance, the AI agent can be maneuvered forward, backward, or rotated left or right. Each execution of these actions provides the agent with a new viewpoint for frame acquisition. To procure a depth image, the "Get Depth Image" function of the AI2-THOR environment is employed, returning a depth image that represents the scene's depth from the AI system's current viewpoint. The depth image is a two-dimensional array, with each element representing the depth value of the corresponding pixel. These depth values serve to comprehend the position and shape of objects within the scene.

The paper randomly selects a scene in the experimental environment, and after initializing the intelligent body in the scene, the movement method and the final location and angle were arbitrarily set, and the RGB, Depth and instance labeled images of the scene were captured. The effect of the model was verified, and the results are shown in Figure 9. The text corresponding to the three detection boxes are *"This is a white door in the front. it is at the far end of the wall."*, *"This is a brown box on the desk. It is near the wall. It is near the wall."*, *"This is a door near the wall. It is a white door."* It can be seen that these description results are relatively accurate, and the model has excellent performance in the test results in the simulation environment.

5 Conclusion

In this paper, the problem of scene semantic description for indoor mobile robots is studied. The ScanNet scene data is processed to obtain its RGBD image, and then the corresponding semantic description is obtained based on the RGBD image. After experiments, we know that the proposed algorithm can effectively describe the indoor scene semantically. The use of multimodal information can help the model understand the scene better and improve the accuracy of the model. Compared with direct RGB image recognition, the proposed model obtains better results in three indexes, such as BLEU, CIDEr, and METEOR, and gets better test performance in the AI2-THOR experimental environment. Overall, the proposed method has high practicality and promotion value and can provide more accurate and advanced semantic information for the perception of indoor mobile robots.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

HW: Investigation, Methodology, Writing—original draft. WW: Conceptualization, Data curation, Supervision, Writing—original draft. WL: Data curation, Supervision, Writing—review & editing. HL: Funding acquisition, Supervision, Writing—review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was jointly supported by Grants from the National Key R&D Program of China (No. 2020AAA0108904), Science and Technology Plan of Shenzhen (No. JCYJ20200109140410340), and Shenzhen Fundamental Research Program (No. GXWD20201231165807007-20200807164903001).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., et al. (2018). "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 6077–6086.

Banerjee, S., and Lavie, A. (2005). "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (AnnArbor, MI), 65–72.

Cai, D., Zhao, L., Zhang, J., Sheng, L., and Xu, D. (2022). "3DJCG: a unified framework for joint dense captioning and visual grounding on 3D point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New Orleans, LA), 16464–16473.

Chen, D. Z., Chang, A. X., and Nießner, M. (2020). "ScanRefer: 3D object localization in RGB-D scans using natural language," in *Computer Vision–ECCV 2020: 16th European Conference* (Glasgow: Springer), 202–221.

Chen, D. Z., Hu, R., Chen, X., Nießner, M., and Chang, A. X. (2022). Unit3D: a unified transformer for 3d dense captioning and visual grounding. *arXiv preprint arXiv:2212.00836*. doi: 10.48550/ARXIV.2212.00836

Chen, Z., Gholami, A., Nießner, M., and Chang, A. X. (2021). "Scan2CAP: context-aware dense captioning in RGB-D scans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3193–3203.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., et al. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*. doi: 10.3115/V1/D14-1179

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., and Nießner, M. (2017). "ScanNet: richly-annotated 3D reconstructions of indoor scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 5828–5839.

Deitke, M., Han, W., Herrasti, A., Kembhavi, A., Kolve, E., Mottaghi, R., et al. (2020). "RoboTHOR: an open simulation-to-real embodied AI platform," in *CVPR* (Seattle, WA).

He, K., Zhang, X., Ren, S., and Sun, J. (2015). "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV), 770–778.

Huang, W., Liu, H., and Wan, W. (2020). An online initialization and self-calibration method for stereo visual-inertial odometry. *IEEE Trans. Robot.* 36, 1153–1170. doi: 10.1109/TRO.2019.2959161

Jiao, Y., Chen, S., Jie, Z., Chen, J., Ma, L., and Jiang, Y.-G. (2022). "MORE: multi-order relation mining for dense captioning in 3D scenes," in *Computer Vision–ECCV 2022: 17th European Conference* (Tel Aviv: Springer), 528–545.

Johnson, J., Karpathy, A., and Fei-Fei, L. (2016). "DenseCap: fully convolutional localization networks for dense captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 4565–4574.

Kannala, J., and Brandt, S. S. (2006). A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. *IEEE Trans. Pattern Anal. Mach. Intell.* 28, 1335–1340. doi: 10.1109/TPAMI.2006.153

Kim, D.-J., Choi, J., Oh, T.-H., and Kweon, I. S. (2019). "Dense relational captioning: triple-stream networks for relationship-based captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA), 6271–6280.

Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., et al. (2017). AI2-THOR: an interactive 3d environment for visual AI. *arXiv preprint arXiv:1712.05474*. doi: 10.48550/arXiv.1712.05474

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Li, W., Liu, H., Tang, H., Wang, P., and Van Gool, L. (2022). "MHFormer: multi-hypothesis transformer for 3D human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, LA), 13147–13156.

Lin, C.-Y. (2004). "ROUGE: a package for automatic evaluation of summaries," in *Proceedings of the Workshop on Text Summarization Branches Out*. 74–81.

Liu, H., Qiu, J., and Huang, W. (2022). "Integrating point and line features for visual-inertial initialization," in *2022 International Conference on Robotics and Automation (ICRA)* (Philadelphia, PA: IEEE), 9470–9476.

Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., and Yuille, A. (2014). Deep captioning with multimodal recurrent neural networks (m-RNN). *arXiv preprint arXiv:1412.6632*. San Diego, CA.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (Philadelphia, PA), 311–318.

Qi, C. R., Yi, L., Su, H., and Guibas, L. J. (2017). "PointNet++: deep hierarchical feature learning on point sets in a metric space," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, eds I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett (Long Beach, CA).

Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.-Y., Johnson, J., et al. (2020). Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501*. doi: 10.48550/arXiv.2007.08501

Ren, S., He, K., Girshick, R. B., and Sun, J. (2015). Faster r-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 1137–1149. doi: 10.1109/TPAMI.2016.2577031

Rubio, F., Valero, F., and Llopis-Albert, C. (2019). A review of mobile robots: concepts, methods, theoretical framework, and applications. *Int. J. Adv. Robot. Syst.* 16, 1729881419839596. doi: 10.1177/17298814198 39596

Shao, Z., Han, J., Marnerides, D., and Debattista, K. (2022). Region-object relation-aware dense captioning via transformer. *IEEE Trans. Neural Netw. Learn. Syst.* doi: 10.1109/TNNLS.2022.31 52990

Sheridan, T. B. (2016). Human–robot interaction: status and challenges. *Hum. Fact.* 58, 525–532. doi: 10.1177/0018720816644364

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). "CIDER: consensus-based image description evaluation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA), 4566–4575.

Wang, H., Zhang, C., Yu, J., and Cai, W. (2022). Spatiality-guided transformer for 3d dense captioning on point clouds. *arXiv preprint arXiv:2204.10688*. doi: 10.24963/IJCAI.2022/194

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., et al. (2015). "Show, attend and tell: neural image caption generation with visual attention," in *International Conference on Machine Learning* (Lille: PMLR), 2048–2057.

Yang, L., Tang, K., Yang, J., and Li, L.-J. (2017). "Dense captioning with joint inference and visual context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 2193–2202.

Yuan, Z., Yan, X., Liao, Y., Guo, Y., Li, G., Cui, S., et al. (2022). "X-Trans2Cap: cross-modal knowledge transfer using transformer for 3d dense captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New Orleans, LA), 8563–8573.

Zhong, Y., Xu, L., Luo, J., and Ma, L. (2022). Contextual modeling for 3d dense captioning on point clouds. *arXiv preprint arXiv:2210.03925*. doi: 10.48550/ARXIV.2210.03925

TYPE Original Research
PUBLISHED 21 November 2023
DOI 10.3389/fnbot.2023.1322645

OPEN ACCESS

EDITED BY
Long Jin,
Lanzhou University, China

REVIEWED BY
Zhongbo Sun,
Changchun University of Technology, China
Liangming Chen,
Chinese Academy of Sciences (CAS), China

*CORRESPONDENCE
Rongbo Lu
✉ lurongbo8563@163.com
Liang Luo
✉ luoliang991105@163.com

RECEIVED 16 October 2023
ACCEPTED 31 October 2023
PUBLISHED 21 November 2023

CITATION
Lu R, Luo L and Liao B (2023) Voting based double-weighted deterministic extreme learning machine model and its application. *Front. Neurorobot.* 17:1322645.
doi: 10.3389/fnbot.2023.1322645

Voting based double-weighted deterministic extreme learning machine model and its application

Rongbo Lu[1]*, Liang Luo[2]* and Bolin Liao[2]

[1]College of Computer and Artificial Intelligence, Huaihua University, Huaihua, China, [2]College of Computer Science and Engineering, Jishou University, Jishou, China

This study introduces an intelligent learning model for classification tasks, termed the voting-based Double Pseudo-inverse Extreme Learning Machine (V-DPELM) model. Because the traditional method is affected by the weight of input layer and the bias of hidden layer, the number of hidden layer neurons is too large and the model performance is unstable. The V-DPELM model proposed in this paper can greatly alleviate the limitations of traditional models because of its direct determination of weight structure and voting mechanism strategy. Through extensive simulations on various real-world classification datasets, we observe a marked improvement in classification accuracy when comparing the V-DPELM algorithm to traditional V-ELM methods. Notably, when used for machine recognition classification of breast tumors, the V-DPELM method demonstrates superior classification accuracy, positioning it as a valuable tool in machine-assisted breast tumor diagnosis models.

KEYWORDS

intelligent learning model, neural network, machine recognition classification, weights determination, machine-assisted diagnosis

1 Introduction

Extreme Learning Machine (ELM) (Huang et al., 2004) is a powerful machine learning algorithm that has emerged as a popular alternative to traditional neural networks [such as Back-Propagation (Haykin, 1998) algorithm (BP) and Levenberg Marquardt (Levenberg, 1944; Marquardt, 1963) algorithm] due to its speed, simplicity, and high performance. ELM is a single-layer feedforward neural network that uses random weight initialization and least-squares optimization to learn from input data (Huang et al., 2006). The algorithm has shown remarkable results in a wide range of applications, from image recognition (Tang et al., 2015) and speech processing (Han et al., 2014) to financial forecasting (Fernández et al., 2019) and anomaly detection (Huang et al., 2015).

One drawback of the ELM algorithm is that the learning parameters of the hidden nodes are randomly assigned and remain unchanged during training, which may lead to a significant impact on its predictive performance and algorithm stability (Gao and Jiang, 2012; Lu et al., 2014). ELM might misclassify certain samples, particularly those near the classification boundaries. In an attempt to address this issue, Cao et al. (2012) proposed a voting-based variant of ELM, referred to as V-ELM. The main idea behind V-ELM is to perform multiple independent ELM trainings instead of a single training, and then make the final decision based on majority voting. However, this approach does not fundamentally resolve the problem of random determination of ELM's various parameters.

Zhang et al. (2014) have highlighted that the performance of Extreme Learning Machine (ELM) is not always optimal when the input weights and hidden layer biases are chosen entirely at random. This randomness is also a significant factor contributing to the redundancy of neurons in the hidden layer of the ELM algorithm (Zhu et al., 2005). In response, scholars have proposed the use of swarm intelligence optimization (Lahoz et al., 2013; Figueiredo and Ludermir, 2014; Zhang et al., 2016), pruning methods (Miche et al., 2009, 2011), and adaptive algorithms (Pratama et al., 2016; Zhao et al., 2017) to optimize the ELM algorithm and enhance its overall performance. However, in practical applications, although these algorithms do succeed in optimizing the number of hidden layer neurons, they introduce a plethora of hyperparameters that typically require iterative optimization, thereby increasing the computational complexity of the algorithm and rendering it challenging to address real-time problems with high time constraints. To tackle this issue, this paper presents an improved algorithm known as Voting based double Pseudo-inverse weights determination Extreme Learning Machine (V-DPELM). The core concept of V-DPELM lies in the stochastic determination of output weights, while input weights are obtained through pseudoinverse calculations. Subsequently, the pseudo-inverse method is employed again to determine optimal output weights, ensuring that both input and output weights are optimal. The obtained DPELM algorithm is subjected to multiple independent trainings, and the final decision is made based on majority voting.

In the 21st century, breast cancer is increasingly recognized as a significant factor negatively impacting the overall quality of life for women worldwide. According to statistics from the World Health Organization (WHO), approximately 1.5 million women suffer greatly from the torment of breast cancer, with approximately 500,000 losing their lives to this disease (Fahad Ullah, 2019). The incidence and mortality rates of breast cancer exhibit a clear and alarming upward trend each year. Research has demonstrated the paramount importance of timely detection, diagnosis, and initiation of treatment in achieving favorable therapeutic outcomes for breast cancer (Lee et al., 2019; Aldhaeebi et al., 2020). Ten crucial features, including symmetry and fractal dimension of breast tumor lesions, play a vital role in determining the nature of the tumor, whether benign or malignant (Wang et al., 2016, 2019). Therefore, it is possible to extract relevant features closely associated with tumor characteristics from acquired patient samples. By employing the proposed V-DPELM algorithm for parameter optimization and subsequent breast tumor classification, the obtained classification and identification results can provide valuable references, assisting physicians in making diagnostic decisions and offering more accurate and rational assessments of patients' conditions.

2 V-DPELM algorithm design

In the section, we first review the basic concept of the traditional ELM algorithm in Section 2.1. Then, we analyzed the DPELM algorithm in Section 2.2. Finally, the new proposed V-DPELM algorithm will be presented in Section 2.3.

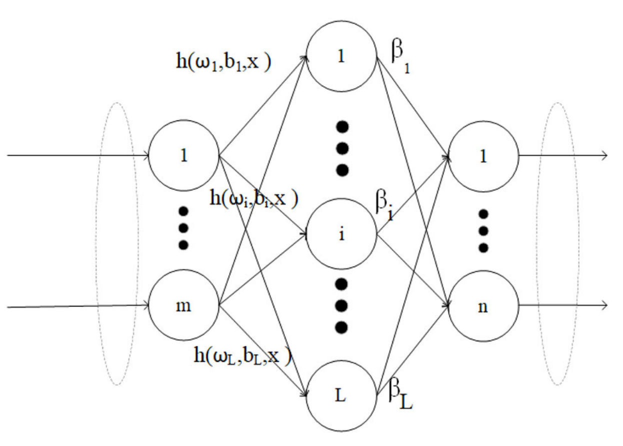


FIGURE 1
ELM network structure.

2.1 Brief review of ELM

Extreme Learning Machine (ELM) is suitable for generalized Single Hidden Layer Feedforward Networks (SLFN). The structure of traditional ELM is similar to SLFN, consisting of three layers: input layer, hidden layer, and output layer. The essence of ELM is that it does not require tuning the hidden layer of SLFN. The structure of ELM is shown in Figure 1.

In the context of N arbitrary training samples $\{(\mathbf{x}_i, \mathbf{t}_i)\}_{i=1}^{N}$, where each sample $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{in})^{\mathrm{T}} \in \mathbb{R}^n$, $\mathbf{t}_i = (t_{i1}, t_{i2}, ..., t_{im})^{\mathrm{T}} \in \mathbb{R}^m$, the resulting output of the ELM with L hidden nodes can be expressed as follows:

$$\mathbf{t}_i = \sum_{j=1}^{L} \boldsymbol{\beta}_j h(\boldsymbol{\omega}_j, b_j, \mathbf{x}_i), i = 1, 2, ..., N \quad (1)$$

Here, $\boldsymbol{\omega_j} = (\omega_{j1}, \omega_{j2}, ..., \omega_{jn})$ represents the weight vector of the jth neuron in the input layer, and b_j is the bias associated with the jth neuron. $h(.)$ indicates the activation function. Furthermore, $\boldsymbol{\beta_j}$ denotes the linked weights between the jth hidden neurons and output neurons, $\boldsymbol{\beta_j} = (\beta_{j1}, \beta_{j2}, ..., \beta_{jm})$.

For all N samples, the equivalent canonical form of linear equation (1) can be expressed as:

$$\mathbf{H}\boldsymbol{\beta} = \mathbf{T}, \quad (2)$$

In Equation (2), T represents the desired output matrix for the training samples, and

$$\mathbf{H} = \begin{bmatrix} h(\boldsymbol{\omega}_1, b_1, \mathbf{x}_1) & \dots & h(\boldsymbol{\omega}_L, b_L, \mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ h(\boldsymbol{\omega}_1, b_1, \mathbf{x}_N) & \dots & h(\boldsymbol{\omega}_L, b_L, \mathbf{x}_N) \end{bmatrix}$$

is the randomized matrix mapping. It is worth noting that the parameters $(\boldsymbol{\omega}_j, b_j)$ of the hidden layer neurons are randomly generated and remain fixed throughout the entire training process of ELM.

The ELM algorithm can be summarized as three steps as follow.

- Step 1: Randomly generate parameters for the hidden layer nodes.
- Step 2: Calculate the output matrix H of the hidden layer.
- Step 3: Calculate the output weight using $\tilde{\boldsymbol{\beta}} = \mathbf{H}^{\dagger}\mathbf{T}$, † represents the pseudo-inverse of the matrix.

2.2 DPELM learning algorithm

Due to the random determination of input weights in traditional ELM, it has resulted in low classification accuracy and an issue of too many hidden layer nodes. Therefore, this section introduces a new method for determining ELM's weights, referred to as the double pseudo-inverse weights determination ELM (DPELM), aiming to enhance its classification accuracy and achieve a more stable structure. DPELM is similar to the traditional ELM network structure, which consists of input layer, hidden layer and output layer. Upon a more comprehensive analysis of the traditional ELM principle, Equation 1 can be reformulated as follows:

$$\mathbf{T} = \boldsymbol{\beta} h(\Omega\mathbf{X} - \mathbf{B}), \tag{3}$$

where $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_N] \in \mathbb{R}^{m\times N}$, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N] \in \mathbb{R}^{n\times N}$, $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_N] \in \mathbb{R}^{L\times N}$, $\boldsymbol{\beta}$ and Ω represent the output weight matrix and the input weight matrix, respectively. Where

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{11} & \beta_{12} & \dots & \beta_{1L} \\ \beta_{21} & \beta_{22} & \dots & \beta_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{m1} & \beta_{m2} & \dots & \beta_{mL} \end{bmatrix} \in \mathbb{R}^{m\times L},$$

$$\boldsymbol{\Omega} = \begin{bmatrix} \omega_{11} & \omega_{12} & \dots & \omega_{1n} \\ \omega_{21} & \omega_{22} & \dots & \omega_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{L1} & \omega_{L2} & \dots & \omega_{Ln} \end{bmatrix} \in \mathbb{R}^{L\times n}.$$

Derivation process: Assuming the bias $\boldsymbol{B}$ and output weight $\boldsymbol{\beta}$ are randomly generated within the interval [a1, a2], and the activation function $h(\cdot)$ is strictly monotonous, the ideal Ω should be equal to $\Omega = (h^{-1}(\boldsymbol{\beta}^{\dagger}\mathbf{T}) + \mathbf{B})\mathbf{X}^{\dagger}$.

Since $\boldsymbol{B}$ and $\boldsymbol{\beta}$ are randomly generated, multiplying both sides of Equation 3 by $\boldsymbol{\beta}^{\dagger}$ results in:

$$\boldsymbol{\beta}^{\dagger}\mathbf{T} = \boldsymbol{\beta}^{\dagger}\boldsymbol{\beta} h(\Omega\mathbf{X} - \mathbf{B}) = h(\Omega\mathbf{X} - \mathbf{B}). \tag{4}$$

By finding the inverse function of the activation function $h(\cdot)$, we can obtain:

$$h^{-1}(\boldsymbol{\beta}^{\dagger}\mathbf{T}) = \Omega\mathbf{X} - \mathbf{B},$$

The above equation can be rewritten as:

$$\Omega\mathbf{X} = h^{-1}(\boldsymbol{\beta}^{\dagger}\mathbf{T}) + \mathbf{B}. \tag{5}$$

Finally, multiplying equation 5 by $\mathbf{X}^{\dagger}$ simultaneously results in

$$\Omega\mathbf{X}\mathbf{X}^{\dagger} = (h^{-1}(\boldsymbol{\Lambda}^{\dagger}\mathbf{Y}) + \Phi)\mathbf{X}^{\dagger},$$

namely,

$$\Omega = (h^{-1}(\boldsymbol{\beta}^{\dagger}\mathbf{T}) + \mathbf{B})\mathbf{X}^{\dagger}.$$

This concludes the proof.

Once the optimal Ω has been determined, the formula $\tilde{\boldsymbol{\beta}} = \mathbf{T}(h(\Omega\mathbf{X} - \mathbf{B}))^{\dagger}$ can be employed to compute the value of $\tilde{\boldsymbol{\beta}}$.

2.3 V-DPELM model training process

Based on theoretical principles, the specific training process for V-DPELM model is outlined as follows:

- Step 1: Given a sample dataset $\aleph = \{(\mathbf{x}_i, \mathbf{t}_i)|\mathbf{x}_i \in \mathbb{R}^n, \mathbf{t}_i \in \mathbb{R}^m\}_{i=1}^N$, where $\mathbf{x}_i$, $\mathbf{t}_i$, N represent the input vector, target vector, and the total number of samples, respectively. This step introduces essential parameters, including the hidden node output function $h(\omega, b, x)$, the count of hidden nodes L, and the number of independent training repetitions K.
- Step 2: Randomly initialize output weights β and hidden layer biases B within the interval [a1, a2].
- Step 3: In the case where the training sample is determined, the optimal input weights Ω are computed using the formula $\Omega = (h^{-1}(\boldsymbol{\beta}^{\dagger}\mathbf{T}) + \mathbf{B})\mathbf{X}^{\dagger}$.
- Step 4: Subsequently, upon obtaining the optimal input weights Ω, the optimal output weights $\tilde{\boldsymbol{\beta}}$ are determined as $\tilde{\boldsymbol{\beta}} = \mathbf{T}(h(\Omega\mathbf{X} - \mathbf{B}))^{\dagger}$.
- Step 5: Repeat steps 2 to 4 for a total of K times to get K independent DPELMs model. Then, perform test tasks on these DPELMs, and the final result is obtained by aggregating the test results using a voting strategy.

The network structure of V-DPELM model is shown in Figure 2. Algorithm 1 provides a specific introduction to the pseudo code of the V-DPELM method.

3 Experimental results and analysis

This section randomly selects 12 datasets from the UCI database to assess the classification performance of the improved Extreme Learning Machine algorithm. All experiments in this paper were conducted using Matlab 2016(a) on a regular PC with an Intel(R) Core(TM) i5-12500H CPU running at 3.60GHz and 16GB of memory.

3.1 Experimental description

The present text conducts a series of experiments to evaluate the performance of the algorithm from various perspectives, including the efficacy of its categorization, the precision of its predictions, the requisite count of neurons within its hidden layers, and the stability of its resultant outputs. The datasets utilized in this research were sourced from the UCI (University of California, Irvine) repository, encompassing both binary classification and multi-classification datasets. It is important to note that the training and test data

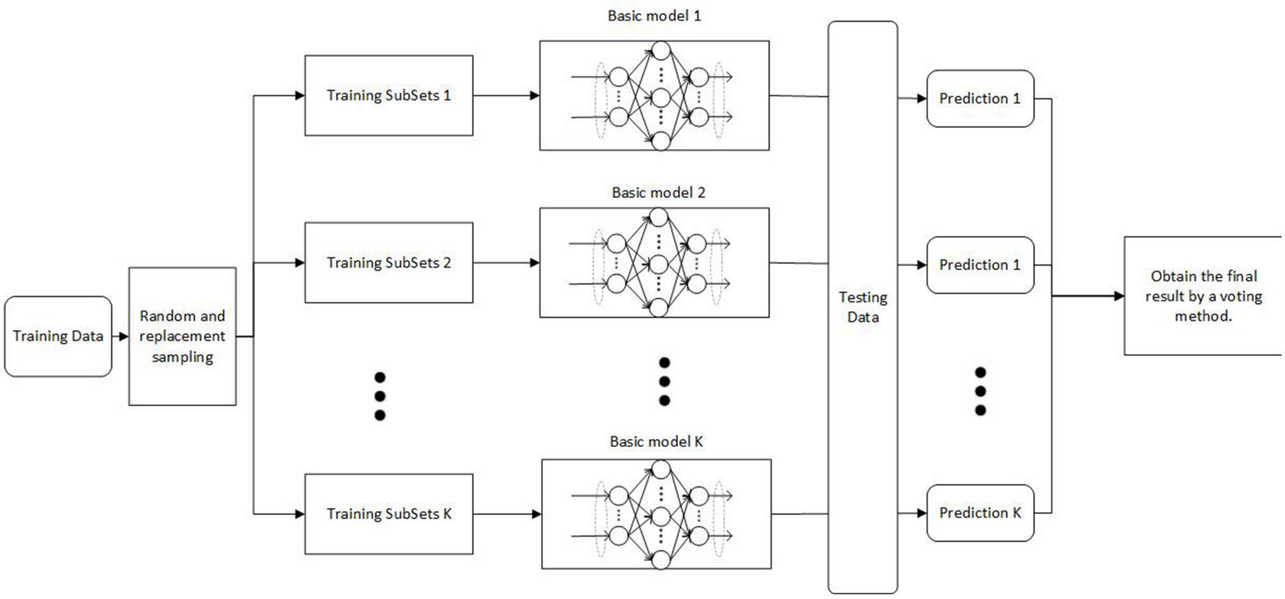


FIGURE 2
V-DPELM network structure.

```
Input: $\aleph = \{(\mathbf{x}_i, \mathbf{t}_i) | \mathbf{x}_i \in \mathbb{R}^n, \mathbf{t}_i \in \mathbb{R}^m\}_{i=1}^N$, hidden active function
    $h(\omega, b, x)$, hidden nodes $L$, independent training
    repetitions $K$, zero valued vector $S_K \in \mathbb{R}^m$;
Output: $TestingAccuracy$;
Set $k = 1$;
while $k \leq K$ do
  Randomly assign the learning parameters $(\beta_i^k, b_i^k)$ of
  the $k$th DPELM;
  Calculate the input weight $\omega^k$;
  Calculate the hidden layer output matrix $H^k$;
  Calculate the output weight $\tilde{\boldsymbol{\beta}}^k$, $\tilde{\boldsymbol{\beta}} = \mathbf{T}(h(\Omega\mathbf{X} - \mathbf{B}))^\dagger$;
  $k = k + 1$;
end while
$c = a + b$;
for all testing sample$x^{test}$ do
  Set $k = 1$;
  while $k \leq K$ do
    using the $k$th trained basic DPELM with leaning
    parameters $(\beta_i^k, b_i^k, \omega_i^k)$ to predict the label of the
    testing sample $x^{test}$;
    Each generated prediction result is then
    stored in $S_K$;
    $k = k + 1$;
  end while
  The final class label of testing sample $x^{test}$ is
  $c^{\mathtt{test}} = \arg\max_{j \in [1, \cdots, m]} \left\{ S_{K, x^{\mathtt{test}}}(j) \right\}$
end for
```

Algorithm 1. V-DPELM.

within each dataset were randomly shuffled for each simulation experiment, ensuring unbiased evaluations. Detailed specifications of these 12 datasets are presented in Table 1.

TABLE 1 Specifications of classification datasets.

Datasets	Attributes	Classes	Samples	Testing data
SL	35	19	215	92
Iris	4	3	100	50
Wine	13	3	100	78
Liver disorders (LD)	6	2	240	105
Pima Indians diabetes (PID)	8	2	537	231
Innosphere	34	2	220	95
Diabetes	8	2	576	191
Balance	4	3	400	225
Ecoli	7	8	100	236
Waveform	21	3	3000	2000
Live	6	2	200	145

3.2 Experimental results and analytical discussion

In this subsection, we begin by employing the Iris dataset, the features of which are displayed in Table 1, to ascertain the efficacy of the V-DPELM algorithm. The corresponding outcomes are illustrated through Figures 3–5 and Table 2. Figures 3, 4 depict the graphs of the confusion matrix. Within these figures, the values along the diagonal of the matrix signify the correctly classified samples, whereas those located elsewhere indicate the misclassified samples.

It is evident that V-DPELM exhibits noteworthy proficiency in performing classification tasks, both in testing and training scenarios. Furthermore, as evident from Figure 5, the optimal classification accuracy reaches approximately 99.5% during testing and 98% during training. Notably, Figure 5 unveils a significant observation: the generalization performance of V-DPELM remains stable even with a modest number of hidden-layer neurons.

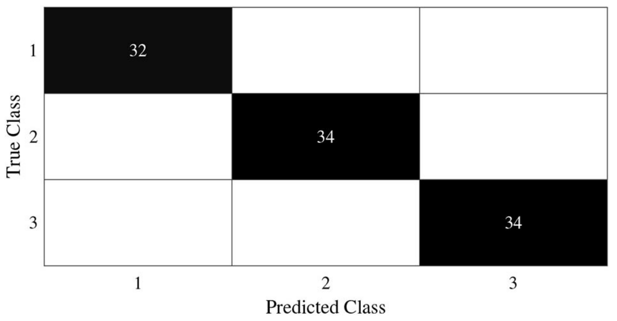


FIGURE 3
Training confusion matrix of Iris dataset.

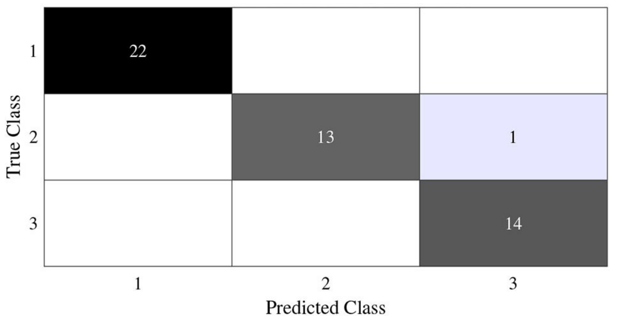


FIGURE 4
Test confusion matrix of Iris dataset.

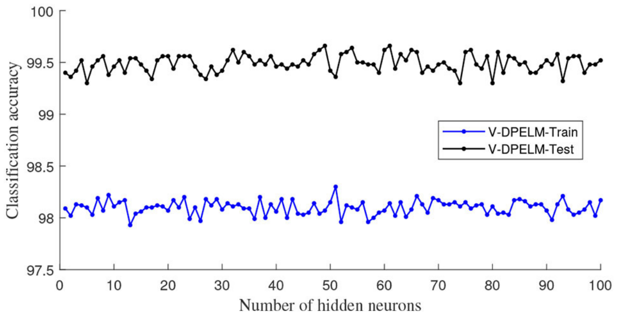


FIGURE 5
V-DPELM classification accuracy for Iris dataset.

TABLE 2 Classification performance of V-DPELM with different hidden layer neuron numbers in the Iris Dataset.

V-DPELM	Accuracy rate (%)		Neurons
	Training	Testing	
	98.09	99.40	1
	98.02	99.36	2
	98.13	99.42	3
	98.12	99.52	4
	98.10	99.30	5
	98.11	99.46	10
	98.07	99.56	20
	98.15	99.42	50
	98.17	99.52	100

TABLE 3 Comparisons of classification accuracy and number of hidden layer neurons of different algorithms.

Datasets	Testing (%)		Hidden layer neurons	
	V-ELM	V-DPELM	V-ELM	V-DPELM
SL	90.25	**92.30**	83	63
Iris	98.42	**99.56**	15	9
Wine	99.38	**99.93**	30	10
Liver Disorders (LD)	73.24	**73.33**	24	7
Pima Indians Diabetes (PID)	81.07	**83.37**	35	30
Innosphere	91.35	**92.88**	47	5
Diabetes	70.96	**81.23**	40	5
Zoo	96.61	**98.22**	20	10
Balance	90.49	**92.08**	40	30
Ecoli	85.23	**89.15**	20	10
Waveform	76.37	**78.31**	80	30
Liver	71.56	**73.79**	20	10

Bold values indicate the maximum value.

This finding is corroborated by Table 2. Specifically, when the count of hidden-layer neurons is set to 3, optimal and consistent classification accuracy is achieved. This phenomenon holds true for other cases as well.

Regarding Table 2, there is an additional aspect that requires elucidation. In the context of assessing the presented growth methodology, the number of hidden-layer neurons in V-DPELM is tuned either manually, with an increment of 1, or automatically

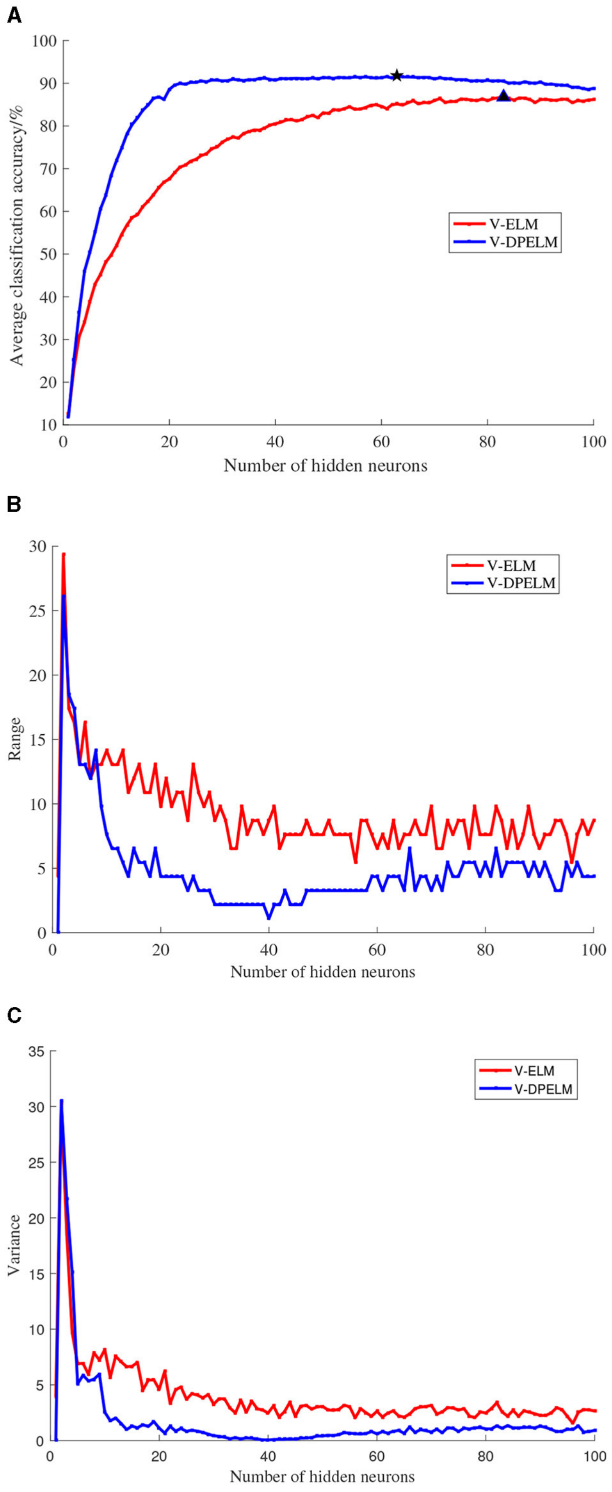


FIGURE 6
SL data set comparison experiment results. **(A)** Changes in classification accuracy. **(B)** Changes in range. **(C)** Changes in variance.

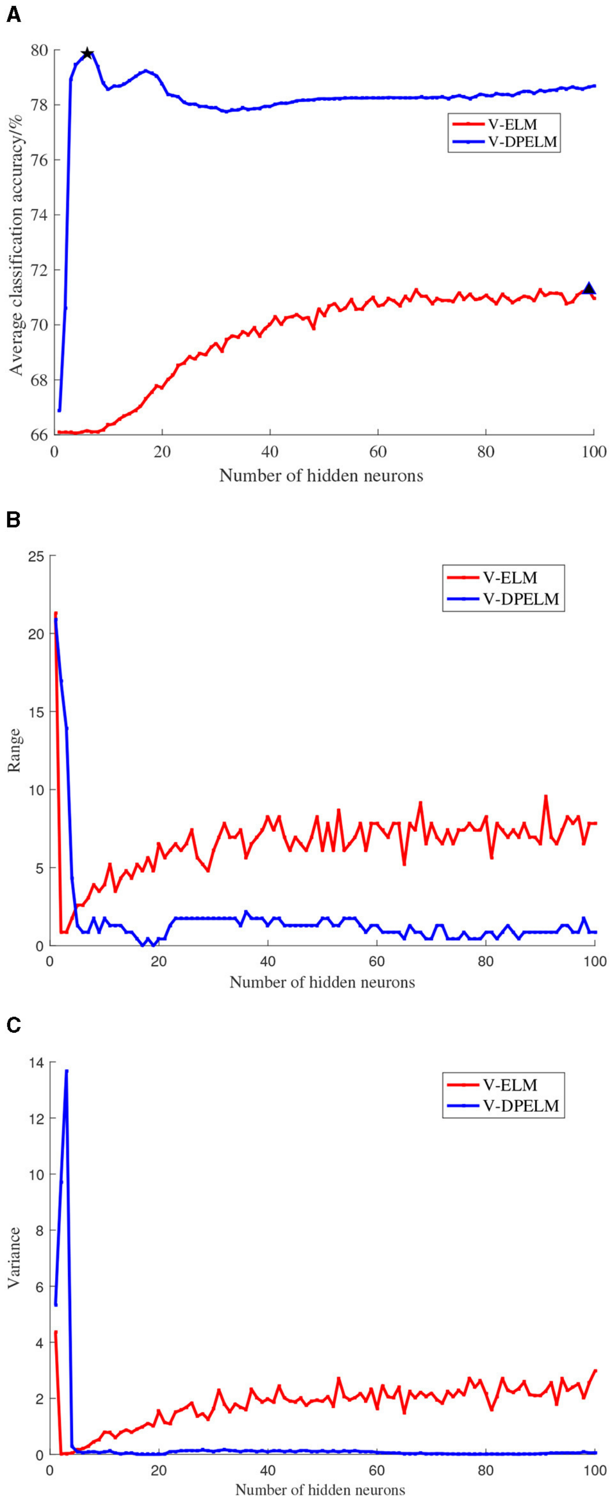


FIGURE 7
Diabetes data set comparison experiment results. **(A)** Changes in classification accuracy. **(B)** Changes in range. **(C)** Changes in variance.

through the growth method. As demonstrated in the table, the proposed growth method effectively identifies the optimal structure for V-DPELM. Consequently, the effectiveness of V-DPELM in pattern classification is preliminarily affirmed.

The impact of the number of neurons in the hidden layer on the predictive performance of both the traditional V-ELM and the algorithm proposed in this study is investigated through experimental comparisons. Initially, a subset of samples from each dataset is selected as training and testing data, with the division between them fixed throughout the experiment. The growing method is employed to determine the number of neurons in the hidden layer, where the accuracy is observed after each addition of one neuron. The corresponding algorithm is considered to have the best network structure when the accuracy remains unchanged or the change falls below a predefined threshold. Subsequently, the ELM algorithm and the algorithm proposed in this paper are executed 100 times within the optimized network structure, and the average classification accuracy is computed using the test dataset. In this experiment, the tangent function (tan) is chosen as the activation function, with its inverse function being the arctangent function (arctan). The comparative analysis of classification accuracy for different algorithms and the required number of neurons in the hidden layer to achieve the highest classification accuracy are presented in Table 3.

From Table 3, it can be observed that the algorithm proposed in this paper outperforms the traditional V-ELM algorithm in terms of classification performance, both in binary datasets and multi-classification datasets. The proposed algorithm achieves higher classification accuracy with fewer neurons in the hidden layer, resulting in a simpler network structure. This indicates that the analytical weight initialization method employed in this paper yields superior results compared to the random weight initialization method. Furthermore, to further analyze the impact of algorithm parameters on classification performance and algorithm stability, this study selects one dataset each from binary and multi-class problems for performance comparison.

The SL dataset, a multi-class dataset, and the Diabetes dataset, a binary classification dataset, are selected for this study. The training and testing sets for both datasets are fixed and unchanged throughout the experiments. The number of neurons in the hidden layer is set to increment from 1 to 100. For each additional neuron, the ELM algorithm and the algorithm proposed in this paper are executed 100 times. The experimental results are analyzed in terms of the mean, variance, and range, as depicted in Figures 6, 7. In these figures, the positions indicated by black pentagons and triangles represent the locations where each algorithm achieves the highest classification accuracy.

Observing Figures 6A, 7A, it becomes evident that the increase in the number of neurons in the hidden layer leads to an initial rapid rise in prediction accuracy for both the traditional V-ELM algorithm and the algorithm proposed in this paper. However, after reaching a certain point, the accuracy levels off or slightly declines. By considering the experimental findings and the Theorem presented in Huang et al. (2006), it can be deduced that the algorithm proposed in this study shares similar characteristics with the traditional V-ELM algorithm. Specifically, as the number of neurons in the hidden layer increases, the algorithm's fitting performance improves. Nevertheless, beyond a critical threshold, further augmenting the number of hidden neurons may cause overfitting on the training samples, resulting in a slower or even decreasing classification accuracy on the test samples.

Furthermore, a thorough examination of Figures 6, 7 reveals that, in both the multi-class SL dataset and the binary Diabetes dataset, the proposed algorithm demonstrates a faster rate of average classification accuracy improvement compared to the conventional V-ELM algorithm. Remarkably, achieving this progress requires a smaller number of neurons in the hidden layer. Additionally, the analysis of variance and range reveals that the proposed algorithm exhibits lower values for both metrics compared to the traditional V-ELM algorithm on the SL and Diabetes datasets. This finding suggests that the proposed algorithm possesses superior stability in comparison to the traditional V-ELM algorithm.

4 Application of V-DPELM in the diagnosis of breast tumors

In order to further validate the accuracy of voting based double pseudo-inverse weights determination extreme learning machine algorithm, this study applies it to the classification and recognition of breast tumor diagnosis. Multiple distinct algorithms are employed to train and recognize the same breast tumor training and testing sets, which are then compared against the performance of the method proposed in this paper.

TABLE 4 Performance comparison of multiple algorithms.

Algorithm	Average classification accuracy (%)	Benign diagnosis rate (%)	Malignant diagnosis rate (%)
V-DPELM	**98.32**	98.67	**97.73**
V-ELM	97.47	**99.93**	93.29
ELM	96.47	96.22	90.13
AFSA-ELM	96.59	96.38	90.61
LVQ	91.57	94.82	85.08
BP	85.88	84.87	88.93

Bold values indicate the maximum value.

4.1 Experimental data

Data in this study were collected from an open data set published by the University of Wisconsin School of Medicine, including 569 cases of breast tumors, 357 benign and 212 malignant. In this paper, 450 groups of tumor data (282 benign cases, 168 malignant cases) were randomly selected as the training set, and the remaining 119 groups of tumor data (75 benign cases, 44 malignant cases) were selected as the test set. Each sample was composed of 30 data, including the mean, standard deviation and maximum value of 10 characteristic values extracted from the breast tumor sample data.

4.2 Experimental results and analysis

For the purpose of comparing algorithmic performance, three performance metrics were considered: the mean diagnostic rate for benign tumors (referred to as benign diagnosis rate), the mean diagnostic rate for malignant tumors (referred to as malignant diagnosis rate), and the average diagnostic accuracy rate. To ensure robustness of the comparison, independent experiments were conducted 20 times for each algorithm, including the proposed algorithm, V-ELM, Artificial Fish Swarm Algorithm-Extreme Learning Machine (AFSA-ELM), ELM, Learning Vector Quantization (LVQ), and Backpropagation Algorithm (BP). The average values of the benign diagnosis rate, malignant diagnosis rate, and overall accuracy rate were calculated and compared. It should be noted that the experimental results for V-ELM, AFSA-ELM, ELM, LVQ, and BP algorithms were sourced from Zhou and Yuan (2017). The comparative findings are summarized in Table 4.

From the findings presented in Table 4, it is apparent that the average accuracy rate achieved by the proposed algorithm surpasses that of the other algorithms. Although the benign diagnosis rate is slightly lower than that of the V-ELM algorithm, the malignant tumor diagnosis rate is considerably higher. These results highlight the efficacy of the proposed algorithm in rapidly and accurately identifying malignant tumors, thus mitigating the risks associated with delayed treatment and potential impacts on treatment efficacy resulting from misdiagnosis.

5 Conclusions

In the 12 randomly selected UCI datasets, the algorithm proposed in this paper, voting based double pseudo-inverse weights determination extreme learning machine algorithm, exhibits varying degrees of improvement in classification performance compared to the traditional V-ELM algorithm. Among these datasets, the Diabetes dataset shows the greatest increase in classification accuracy, with a significant enhancement of 10.27%. On the other hand, the LD dataset demonstrates the smallest improvement, with a marginal increase of only 0.09% in classification accuracy.

Moreover, the improved algorithm achieves optimal classification accuracy with fewer hidden layer neurons compared to the traditional ELM algorithm, resulting in a simpler network structure.

Additionally, the improved algorithm exhibits reduced variance and range in both the SL and Diabetes dataset experiments, indicating enhanced stability. Furthermore, in the breast tumor classification and recognition experiments, the diagnostic performance of the proposed algorithm surpasses that of V-ELM, AFSA-ELM, ELM, LVQ, and BP methods. This observation highlights the advantage of the proposed algorithm in achieving high classification accuracy in breast tumor auxiliary diagnosis. Thus, the application of this method for breast tumor auxiliary diagnosis is deemed feasible. In addition, it is worth pointing out that processing multi-dimensional data can be a research direction for future work.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://archive.ics.uci.edu/datasets.

Author contributions

RL: Funding acquisition, Investigation, Supervision, Validation, Writing—review & editing. LL: Conceptualization, Data curation, Formal analysis, Project administration, Resources, Software, Visualization, Writing—original draft, Writing—review & editing. BL: Investigation, Methodology, Writing—review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported in part by the National Natural Science Foundation of China under Grant No. 62066015.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aldhaeebi, M. A., Alzoubi, K., Almoneef, T. S., Bamatraf, S. M., Attia, H., and Ramahi, O. M. (2020). Review of microwaves techniques for breast cancer detection. *Sensors* 20, 2390. doi: 10.3390/s20082390

Cao, J., Lin, Z., Huang, G.-B., and Liu, N. (2012). Voting based extreme learning machine. *Inf. Sci.* 185, 66–77. doi: 10.1016/j.ins.2011.09.015

Fahad Ullah, M. (2019). Breast cancer: current perspectives on the disease status. *Adv. Exp. Med. Biol.* 1152, 51–64. doi: 10.1007/978-3-030-20301-6_4

Fernández, C., Salinas, L., and Torres, C. E. (2019). A meta extreme learning machine method for forecasting financial time series. *Appl. Intell.* 49, 532–554. doi: 10.1007/s10489-018-1282-3

Figueiredo, E. M., and Ludermir, T. B. (2014). Investigating the use of alternative topologies on performance of the pso-elm. *Neurocomputing* 127, 4–12. doi: 10.1016/j.neucom.2013.05.047

Gao, G.-Y., and Jiang, G.-P. (2012). Prediction of multivariable chaotic time series using optimized extreme learning machine. *Acta Phys. Sin.* 61, 040506. doi: 10.7498/aps.61.040506

Han, K., Yu, D., and Tashev, I. (2014). "Speech emotion recognition using deep neural network and extreme learning machine, in *In Interspeech 2014*. doi: 10.21437/Interspeech.2014-57

Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation*. Hoboken, NJ: Prentice Hall PTR.

Huang, G.-B., Bai, Z., Kasun, L. L. C., and Vong, C. M. (2015). Local receptive fields based extreme learning machine. *IEEE Comput. Intell. Magaz.* 10, 18–29. doi: 10.1109/MCI.2015.2405316

Huang, G.-B., Zhu, Q.-Y., and Siew, C.-K. (2004). "Extreme learning machine: a new learning scheme of feedforward neural networks, in *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)* (IEEE), 985–990.

Huang, G.-B., Zhu, Q.-Y., and Siew, C.-K. (2006). Extreme learning machine: theory and applications. *Neurocomputing* 70, 489–501. doi: 10.1016/j.neucom.2005.12.126

Lahoz, D., Lacruz, B., and Mateo, P. M. (2013). A multi-objective micro genetic elm algorithm. *Neurocomputing* 111, 90–103. doi: 10.1016/j.neucom.2012.11.035

Lee, K., Kruper, L., Dieli-Conwright, C. M., and Mortimer, J. E. (2019). The impact of obesity on breast cancer diagnosis and treatment. *Curr. Oncol. Rep.* 21, 1–6. doi: 10.1007/s11912-019-0787-1

Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quart. Appl. Mathem.* 2, 164–168. doi: 10.1090/qam/10666

Lu, H. J., An, C. L., Zheng, E. H., and Lu, Y. (2014). Dissimilarity based ensemble of extreme learning machine for gene expression data classification. *Neurocomputing* 128, 22–30. doi: 10.1016/j.neucom.2013.02.052

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Ind. Appl. Mathem.* 11, 431–441. doi: 10.1137/0111030

Miche, Y., Sorjamaa, A., Bas, P., Simula, O., Jutten, C., and Lendasse, A. (2009). OP-ELM: optimally pruned extreme learning machine. *IEEE Trans. Neural Netw.* 21, 158–162. doi: 10.1109/TNN.2009.2036259

Miche, Y., Van Heeswijk, M., Bas, P., Simula, O., and Lendasse, A. (2011). Trop-elm: a double-regularized elm using lars and tikhonov regularization. *Neurocomputing* 74, 2413–2421. doi: 10.1016/j.neucom.2010.12.042

Pratama, M., Zhang, G., Er, M. J., and Anavatti, S. (2016). An incremental type-2 meta-cognitive extreme learning machine. *IEEE Trans. Cybern.* 47, 339–353. doi: 10.1109/TCYB.2016.2514537

Tang, J., Deng, C., and Huang, G.-B. (2015). Extreme learning machine for multilayer perceptron. *IEEE Trans. Neural Netw. Learn. Syst.* 27, 809–821. doi: 10.1109/TNNLS.2015.2424995

Wang, Z., Li, M., Wang, H., Jiang, H., Yao, Y., Zhang, H., et al. (2019). Breast cancer detection using extreme learning machine based on feature fusion with cnn deep features. *IEEE Access* 7, 105146–105158. doi: 10.1109/ACCESS.2019.2892795

Wang, Z., Qu, Q., Yu, G., and Kang, Y. (2016). Breast tumor detection in double views mammography based on extreme learning machine. *Neural Comput. Applic.* 27, 227–240. doi: 10.1007/s00521-014-1764-0

Zhang, W.-B., Ji, H.-B., Wang, L., and Zhu, M.-Z. (2014). Multiple hidden layer output matrices extreme learning machine. *Syst. Eng. Electr*. 36, 1656–1659.

Zhang, Y., Wu, J., Cai, Z., Zhang, P., and Chen, L. (2016). Memetic extreme learning machine. *Patt. Recogn.* 58, 135–148. doi: 10.1016/j.patcog.2016.04.003

Zhao, Y.-P., Li, Z.-Q., Xi, P.-P., Liang, D., Sun, L., and Chen, T.-H. (2017). Gram-schmidt process based incremental extreme learning machine. *Neurocomputing* 241, 1–17. doi: 10.1016/j.neucom.2017.01.049

Zhou, H.-,p. and Yuan, Y. (2017). Application of elm in computer-aided diagnosis of breast tumors based on improved fish swarm optimization algorithm. *Comput. Eng. Sci.* 39, 2145.

Zhu, Q.-Y., Qin, A. K., Suganthan, P. N., and Huang, G.-B. (2005). Evolutionary extreme learning machine. *Patt. Recogn.* 38, 1759–1763. doi: 10.1016/j.patcog.2005.03.028

TYPE Original Research
PUBLISHED 05 January 2024
DOI 10.3389/fnbot.2023.1294211

Check for updates

OPEN ACCESS

EDITED BY
Long Jin,
Lanzhou University, China

REVIEWED BY
Olarik Surinta,
Mahasarakham University, Thailand
Saifullah Tumrani,
Heidelberg University, Germany

*CORRESPONDENCE
Bochao Su
✉ subochao@szpt.edu.cn

RECEIVED 14 September 2023
ACCEPTED 05 December 2023
PUBLISHED 05 January 2024

CITATION
Sun X, Chen Y, Duan Y, Wang Y, Zhang J, Su B and Li L (2024) Vehicle re-identification method based on multi-attribute dense linking network combined with distance control module.
Front. Neurorobot. 17:1294211.
doi: 10.3389/fnbot.2023.1294211

Vehicle re-identification method based on multi-attribute dense linking network combined with distance control module

Xiaoming Sun[1], Yan Chen[1], Yan Duan[1], Yongliang Wang[1], Junkai Zhang[1], Bochao Su[2]* and Li Li[3]

[1]Heilongjiang Province Key Laboratory of Laser Spectroscopy Technology and Application, Harbin University of Science and Technology, Harbin, China, [2]Institute of Intelligent Manufacturing Technology, Shenzhen Polytechnic, Shenzhen, China, [3]School of Mathematics, Harbin Institute of Technology, Harbin, China

Introduction: Vehicle re-identification is a crucial task in intelligent transportation systems, presenting enduring challenges. The primary challenge involves the inefficiency of vehicle re-identification, necessitating substantial time for recognition within extensive datasets. A secondary challenge arises from notable image variations of the same vehicle due to differing shooting angles, lighting conditions, and diverse camera equipment, leading to reduced accuracy. This paper aims to enhance vehicle re-identification performance by proficiently extracting color and category information using a multi-attribute dense connection network, complemented by a distance control module.

Methods: We propose an integrated vehicle re-identification approach that combines a multi-attribute dense connection network with a distance control module. By merging a multi-attribute dense connection network that encompasses vehicle HSV color attributes and type attributes, we improve classification rates. The integration of the distance control module widens inter-class distances, diminishes intra-class distances, and boosts vehicle re-identification accuracy.

Results: To validate the feasibility of our approach, we conducted experiments using multiple vehicle re-identification datasets. We measured various quantitative metrics, including accuracy, mean average precision, and rank-n. Experimental results indicate a significant enhancement in the performance of our method in vehicle re-identification tasks.

Discussion: The findings of this study provide valuable insights into the application of multi-attribute neural networks and deep learning in the field of vehicle re-identification. By effectively extracting color information from the HSV color space and vehicle category information using a multi-attribute dense connection network, coupled with the utilization of a distance control module to process vehicle features, our approach demonstrates improved performance in vehicle re-identification tasks, contributing to the advancement of smart city systems.

KEYWORDS

vehicle re-identification, multi-attributes, HSV color space, dense connection network, distance control module

1 Introduction

To strengthen road traffic management, the coverage rate of urban road traffic monitoring is increasing, resulting in the daily generation of more video image data. As the volume of video data reaches a certain threshold, the deployment of personnel for monitoring and control becomes inadequate. Consequently, vehicle recognition technology has been introduced. Vehicle re-identification aims to identify the same vehicle in different locations and at different times based on the vehicle information collected by a fixed-position sensor.

As early as 1998, Coifman B recalculated features, such as the effective vehicle length between two continuous metric stations on the highway (Coifman, 1998). This method is too limited; it represents an early form of vehicle re-identification. Abdulhai and Tabib (2018) attempted to improve the accuracy of vehicle re-identification at continuous loop detection stations by enhancing the mode proximity distance metric in the pattern recognition process. Relevant experiments were not completed but showed the potential to enhance the accuracy. Liu et al. (2016a) created a vehicle re-identification (VeRi) dataset based on real urban surveillance scenes. Since then, research in re-identification has progressed rapidly. Zhu et al. (2018) proposed a joint deep learning method (JFSDL) for vehicle re-identification. The Siamese Deep Network is used to extract the features of the input vehicle image pairs, and the similarity score between the input vehicle image pairs is obtained based on the hybrid similarity learning function. Lou et al. (2019a) created a new super large vehicle re-identification dataset VERI-wild, which contains more than 400,000 images of 40,000 vehicles. Zheng et al. (2021) used four public vehicle datasets to create a unique large vehicle dataset called VehicleNet and developed a two-step progressive approach to learn more robust visual representations from VehicleNet. Qian et al. (2020) proposed a deep convolutional neural network (SAN) based on dual branching and attribute perception to learn effective feature embedding for vehicle recognition tasks. Ratnesh et al. (2020) used triplet embedding to solve the problem of vehicle re-identification *in camera* networks. Peng et al. (2020) proposed an adaptive vehicle re-identification domain adaptive framework (DAVR) that uses the tag data from the source domain to adapt to the target domain, reducing cross-domain bias. Teng et al. (2020) proposed a multiview branch network, where each branch learns a view-specific feature and introduces a spatial attention model into each feature-learning branch to strengthen the ability to discriminate local differences. Jin et al. (2021) proposed a multicentric metric learning method for vehicle re-identification in multiple views. Zhang et al. (2022) proposed a double attention granularity network (DAG-Net) for vehicle re-identification. The dual-branch neural network was used to extract coarse-grained and fine-grained features, and a self-attention model was added to each branch to enable DAG-Net to recognize different regions of interest (ROIs) at coarse and fine levels for coarse-grained and fine-grained identification. Subsequently, Guo et al. (2019) proposed a novel two-stage attention network supervised by the Top-k Accuracy Multiple Granularity Ranking Loss (TAMR), aiming to learn effective feature embedding for the vehicle re-identification task. Hou et al. (2019) introduced the Deep Quadruplet-wise Adaptive Learning method (DQAL), which introduces the concept of quadruplets and generates four sets of inputs. By combining the proposed quadruplet network loss and softmax loss, they developed a quadruplet network to learn more discriminative vehicle recognition features. Zhang et al. (2019) introduced the Partial Guidance Attention Network (PGAN), effectively integrating global and partial information for discriminative feature learning. Bashir et al. (2019) took the pioneering approach of addressing vehicle re-identification in an unsupervised manner, utilizing a progressive two-step cascaded framework to formulate the entire vehicle re-identification problem as an unsupervised learning paradigm. PAMAL (Tumrani et al., 2020) utilized multi-attribute features, i.e., color and type, and vehicle key points to solve the re-identification task. MSCL (Yuefeng et al., 2022) achieves unsupervised vehicle re-identification through the integration of the Discrete Sample Separation module and Mixed Sample Contrastive Learning. VAAG (Tumrani et al., 2023) addresses the re-identification task by learning robust discriminative features encompassing camera views, vehicle types, and vehicle colors.

In summary, an algorithm for classifying the color features of vehicles based on the HSV color space is proposed. The image is transformed into the HSV color space, and saturation (S) and brightness (V) are introduced, which are sensitive to the reflection coefficient of the object surface. The color features in the HSV color space are extracted by a feature extraction network for accurate color attribute classification. Second, based on the concepts of the YOLO model and DenseNet network, an improved densely connected vehicle classification network is designed by integrating the extracted color features in the HSV color space. The improved network model is used to obtain different dimensional features for the image of the target vehicle, reducing the amount of computation and improving the feature usage rate. The results of the different dimensional features are weighted and fused to improve the accuracy of vehicle classification. It is combined with the vehicle re-identification network to quickly propose class-independent images for the re-identification network. Based on the traditional vehicle recognition network, a new distance control block (DC module) is developed in this study. According to the feature extraction network, the features extracted from the image are processed by similarity DC or difference DC to shorten the feature distance within the image class and increase the feature distance between image classes. Finally, the performance of this algorithm is verified by experiments.

2 Methods

2.1 Multi-attribute dense link classification

In this section, a vehicle classification method based on a dense network with multiple attributes is proposed. The test images are filtered, and the images that are similar to the target vehicle are re-recognized to eliminate the images that do not match the target vehicle class. There are many vehicle attributes, such as model, color, detail features, and volume. In this section, the classification of the dense connection of several attributes is continued, and the most characteristic model and color are selected as the research objects. This method uses a dense connectivity structure to reduce the computational overhead in the network. It combines the color features in the HSV space to minimize the impact of the external environment on vehicle color recognition. The flowchart is shown in Figure 1A. The individual steps are as follows.

2.1.1 Color feature extraction

There are various colors of vehicles. In this study, the colors of vehicles are classified into 10 categories: yellow, orange, green, gray, red, blue, white, gold, brown, and black.

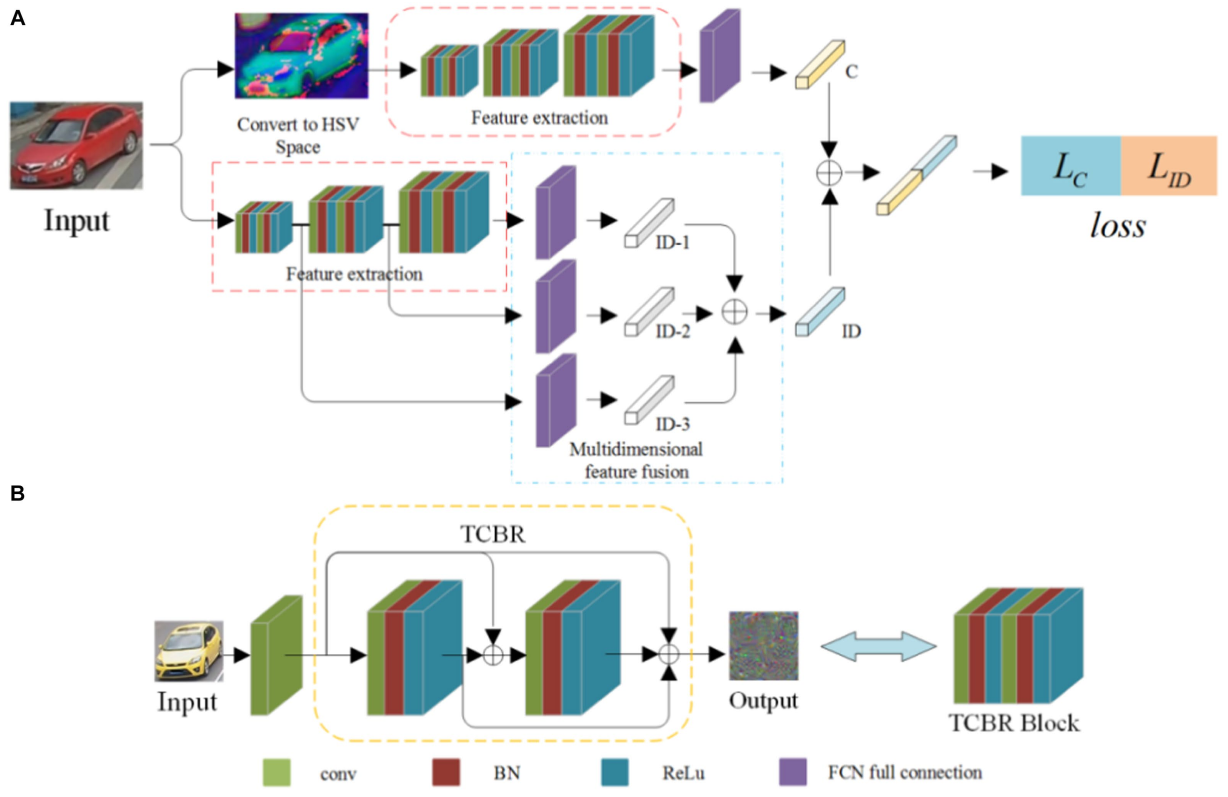


FIGURE 1
Flowchart of the classification with dense connection and multiple attributes. **(A)** Schematic diagram of the multi-attribute dense connection network structure, **(B)** Structural diagram of the small dense connecting block.

(1) Convert the RGB image to an HSV image, as shown in Formula (1).

$$\max(i,j) = \max\left[I_R(i,j), I_G(i,j), I_B(i,j)\right]$$
$$\min(i,j) = \min\left[I_R(i,j), I_G(i,j), I_B(i,j)\right] \quad (1)$$
$$\Delta = \max(i,j) - \min(i,j)$$

Where $I_R(i,j)$, $I_G(i,j)$, and $I_B(i,j)$ represent the values of the R component, G component, and B component corresponding to the pixel coordinate (i,j) points, $\max(i,j)$ represents the maximum value among the R, G, and B components, $\min(i,j)$ represents the minimum value among the R, G, and B components, and $\varnothing$ takes the difference between the two, representing the span of the three components.

The values of the *H* component, *S* component, and *V* component are calculated according to Formula 2. The calculation involves determining the values of the *H* component, *S* component, and *V* component in the HSV color space.

$$H(i,j) = \begin{cases} [I_G(i,j) - I_B(i,j)] / \Delta \times 60, & \max(i,j) = I_R(i,j) \\ 120 + [I_B(i,j) - I_R(i,j)] / \Delta \times 60, & \max(i,j) = I_G(i,j) \\ 240 + [I_R(i,j) - I_G(i,j)] / \Delta \times 60, & \max(i,j) = I_B(i,j) \end{cases}$$
$$V(i,j) = \max(i,j)$$
$$S(i,j) = \Delta / \max(i,j) \quad (2)$$

Where $H(i,j)$, $S(i,j)$, and $V(i,j)$ represent the values of *H* component, *S* component, and *V* component corresponding to the pixel with coordinate (i,j) converted to HSV color space;

(2) Color feature extraction

The structure of the feature extraction network consists of three TCBR blocks. As shown in Figure 1B, each TCBR block is composed of two CBR blocks, and each CBR block is made up of a convolution layer, a BN layer, and a ReLU layer. The TCBR block is a Twice Convolution Batch Normalization ReLU block structure. In the TCBR block, all outputs are summed before the input of the second CBR block, and the features of the input, the first output, and the double output are summed as the input of the next layer, i.e., the dense connections. The outputs of each output node are directly summed, ensuring consistent dimensions for the results of each output node. This reduces the computation of the network, and the dimension conversion is achieved by adding a convolutional layer between two TCBR blocks, making the network more flexible.

The feature extraction network is shown in Figure 1A. Each TCBR block performs a dimensional transformation through the convolutional layer and the pooling layer, extracting features of different dimensions to obtain high-dimensional features of the image.

The extracted high-dimensional features are fed into the fully connected layer, mapping the features to the sample space. Subsequently, the color feature vector *C* corresponding to the HSV features is obtained through regression.

2.1.2 Extraction of the category characteristics

In this study, vehicles are classified into eight categories. The category designations from 1 to 8 are sedan, SUV, van, hatchback, MPV, pickup, bus, and truck. Figure 2 shows schematic representations of these eight vehicle types.

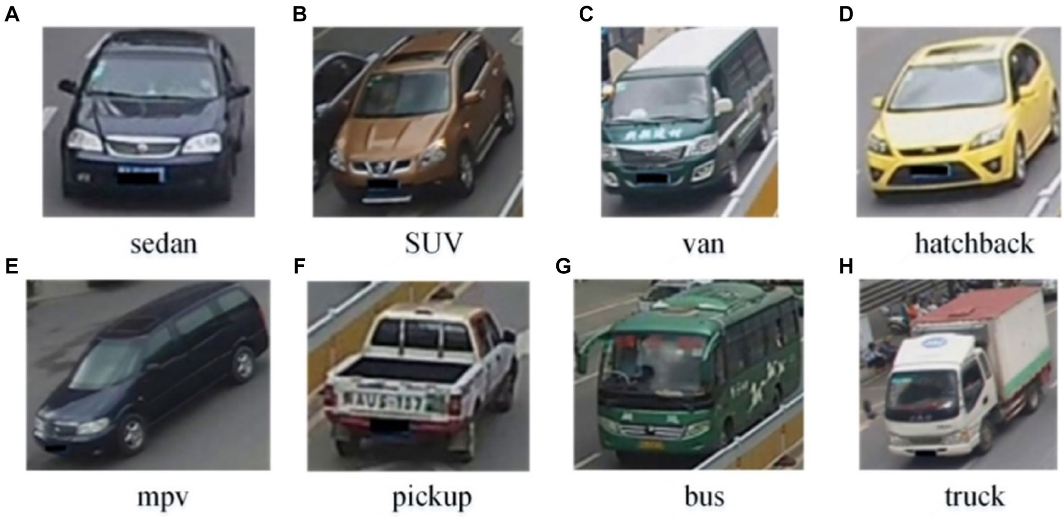


FIGURE 2
Example images of eight vehicle types. **(A)** sedan; **(B)** SUV; **(C)** van; **(D)** hatchback; **(E)** mpv; **(F)** pickup; **(G)** bus; **(H)** truck.

(1) Multidimensional Feature Extraction

In multi-dimensional feature extraction, the TCBR module mentioned above is utilized to extract multi-dimensional features. To better eliminate various features of the vehicle, the feature extraction network is correspondingly improved, as shown in Figure 1. Three different dimensions of features are extracted, namely, $ID-1$, $ID-2$, and $ID-3$.

(2) Multi-dimension feature fusion

According to Formula 3, the three eigenvectors ($ID-1$,$ID-2$, and $ID-3$) are:

$$\begin{aligned} f_1 &= \left(p_A, p_B, p_C, \ldots, p_H\right) \\ f_2 &= \left(p_B, p_A, p_C, \ldots, p_H\right) \\ f_3 &= \left(p_C, p_A, p_B, \ldots, p_H\right) \end{aligned} \tag{3}$$

where p_A ($A-H$ corresponds to eight vehicle types) is the maximum value in f_1, indicating that the probability of the picture being type A is the highest. Similarly, p_B is the maximum value in f_2, signifying that the image has the highest probability of being type B; p_C is the maximum value in f_3, indicating that the image has the highest probability of being class C. The classes A, B, and C are distinguished for better comprehension. In practice, these classes (A, B, and C) can be the same.

Then, the scores are calculated. For type A, as shown in Formula (4).

$$w_{A1} = \frac{f_1\left(p_A\right)}{f_1\left(p_A\right) + f_2\left(p_A\right) + f_3\left(p_A\right)} \tag{4}$$

Where w_{A1} is the weight coefficient of p_A in the weight value of type A in the vector f_1. Similarly, w_{A2} and w_{A3} can be obtained. As shown in Formula (5):

$$\begin{aligned} w_{A2} &= \frac{f_2\left(p_A\right)}{f_1\left(p_A\right) + f_2\left(p_A\right) + f_3\left(p_A\right)} \\ w_{A3} &= \frac{f_3\left(p_A\right)}{f_1\left(p_A\right) + f_2\left(p_A\right) + f_3\left(p_A\right)} \end{aligned} \tag{5}$$

Finally, the score S_A of type A is as shown in Formula (6).

$$S_A = \left[w_{A1}{}^{*} f_1\left(p_A\right) + w_{A2}{}^{*} f_2\left(p_A\right) + w_{A3}{}^{*} f_3\left(p_A\right)\right] \tag{6}$$

As a result, the values S_B and S_C of type B and type C are determined in the same way. One compares three values and takes the highest corresponding type as the classification result.

2.1.3 Multi-attribute dense connection classification

The *output* of the feature classification network is a one-dimensional vector, as shown in Formula (7).

$$output = \left[C, ID\right] \tag{7}$$

where C represents the color feature information in the HSV space of the vehicle in the image, which is a $one-hot\left(10\right)$ vector, representing the normalized value of the ratings corresponding to the 10 color categories; ID represents the class feature information of the vehicle in the image, which is a $one-hot\left(8\right)$ vector, representing the normalized value of the ratings corresponding to the eight vehicle categories.

2.1.4 Loss function

The network receives the color feature information and the vehicle category feature information simultaneously, so the loss function also has two parts, namely, the color feature loss and the vehicle category feature loss. The loss function is developed based on the cross-entropy loss. The loss function for color features is shown in Formula (8).

$$L_C = -\sum_{\mathrm{i}=1}^{n} q\left(i\right)\log\left(p\left(i\right)\right) \tag{8}$$

where n represents the number of color categories and assigns the color attribute to the color attribute category. $p\left(i\right)$ represents the probability that the image belongs to category i, $q\left(i\right)$ is a symbolic function. If category *i* is a basic category, the value is 1, and if category *i* is not a real category, the value is 0.

The loss function L_{ID} of vehicle category characteristic loss is as shown in Formula (9).

$$L_{ID} = -\sum_{j=1}^{m} q(j)\log(p(j)) \tag{9}$$

where *m* is the number of categories of vehicle, which was given before. There are eight types, $p(j)$ denotes the probability that the image belongs to type *j*, and $q(j)$ is also a symbolic function. If the category *j* is an objective type, the value is 1, and if the category *j* is not an objective type, the value is 0.

The final network *loss* function *loss* is shown in Formula (10):

$$loss = L_C + L_{ID} \tag{10}$$

2.2 Vehicle re-identification

After classifying the dense connection of multiple attributes, the system algorithm has filtered out the vehicles with the same color and category as the query vehicle. Then, the final process of re-identifying the vehicle is performed. As shown in Figure 3, the network first extracts features from the input image and obtains high-dimensional features *A*. Then, based on the features of the same category in the feature set, the feature *A* is processed by a similarity DC module and a differential DC module, and the two features are merged into a new feature *A*′.

2.2.1 Feature extraction

The function of the feature extraction network is to extract the high-dimensional features of the input images. To reduce the computational cost, the TCBR module proposed in Chapter 3 is used. As shown in Figure 3, it can be observed that the network consists of two cascaded TCBR modules, and the convolutional layer in the middle is used for dimension conversion. The shape of the input image is set to 224 × 224. First, the convolution layer with the size of the fusion kernel b, the number of fusion kernels 50, and the step size 1 is introduced to perform the pre-convolution. Then, the first TCBR module is instructed to perform dense convolution with multiple inputs in 50 dimensions. Moreover, a convolution layer with the size of the convolution kernel 3 × 3, the number of convolution kernels 100, and the step size 1 is introduced to perform the dimension transformation. As the last step, the second TCBR module is instructed to perform the convolution of features with multiple inputs in 100 dimensions, and then the feature map is output.

2.2.2 Feature set production

The feature set corresponds to the training set used, and one image is selected from each category. Assuming that the total number of classes is the same, the feature extraction network from the previous section is used to extract the features, and the extracted features are integrated into the feature set. The feature set is the feature vector cluster with the category number.

2.2.3 DC module processing

Introduction of DC module: the DC module is divided into two types, one is the similarity module DC, which is based on the target image and uses the comparison image for similarity pooling; the other is the difference module DC, which is based on the target image and performs difference pooling of the contrast image. Each point in the high-dimensional feature space represents the corresponding semantic features of that part. That is, in a sense, they are the domain features. For the similarity module DC, the image after processing attenuates the influence of the prominent features (e.g., the lamp and window position features are identical to the target image).

In contrast, after processing in the DC module, the image may enhance the influence of secondary features (such as body and other parts). For the positive sample (i.e., the image belonging to the same vehicle as the target image), the influence of secondary features is greater than in the negative sample. After two DC modules, the distance between the target image and the positive example is "close." For negative examples, the secondary feature itself is smaller than in positive examples. After processing two DC modules, the influence of secondary features, "pulling away" and distance of the target image, is increased.

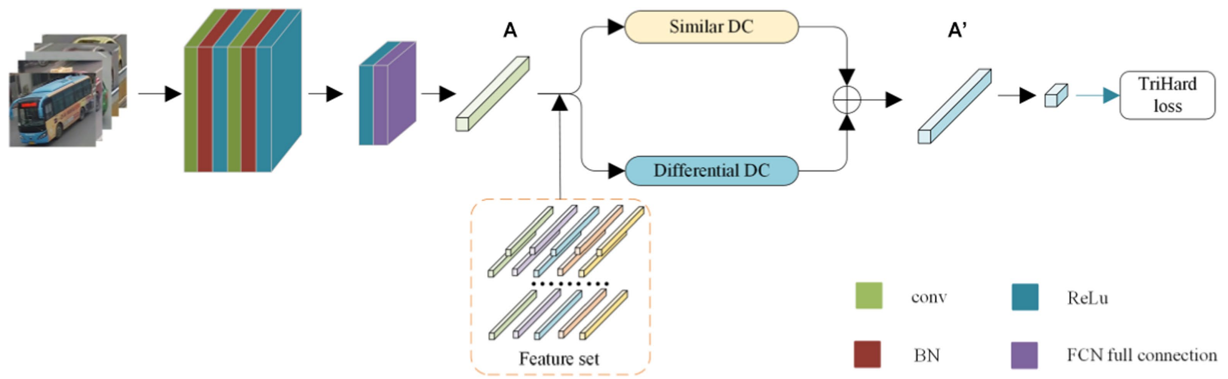


FIGURE 3
Vehicle re-identification model based on DC module.

(1) Similarity DC module schematic diagram is as follows:

As shown in Figure 4, the schematic diagram of a similar DC module is presented. It can be observed from Figure 4 that the core size of the module is 3×3. After extracting the input sample pair, the feature map is traversed by a window of size 3×3, with a step size of 2. The difference value of the corresponding pixels in the window is calculated, and the pixel value corresponding to the minimum difference value is selected to replace the pixel value of the points in the window. The image is divided into a small window of size 3×3. $A_i\left(i=1,2,3,\ldots,9\right)$ is used to represent the values of 9 points in the target image A window, and $B_i\left(i=1,2,3,\ldots,9\right)$ is used to describe the values of 9 points in the contrasting image B window. The distance $D_i\left(i=1,2,3,\ldots,9\right)$ between the corresponding points is calculated, as shown in Formula (11).

$$D_i=\left|A_i-B_i\right|\left(i=1,2,3,\ldots,9\right) \tag{11}$$

The minimum value $D_{\min}$ in D_i is obtained as shown in Formula (12).

$$D_{\min}=\min\left\{D_1,D_2,D_3,\ldots,D_9\right\} \tag{12}$$

The corresponding pixel index value m for $D_{\min}$ is shown in Formula (13).

$$m=\Sigma f\left(i\right)\quad\left(i=1,2,3,\ldots,9\right) \tag{13}$$

The m value obtained by Formula (13) is the index value of the nearest point in the corresponding window between the target image A and the contrast image B, and the $f\left(i\right)$ definitions are as shown in Formula (14).

$$f\left(i\right)=\begin{cases} i, & D_i=D_{\min} \\ 0, & D_i\neq D_{\min}\end{cases} \tag{14}$$

Then, the value of all points is replaced in the contrast image B window with the value B_{m} corresponding to point m, as shown in Formula (15).

$$B_i=B_{\mathrm{m}},\quad\left(i=1,2,3,\ldots,9\right) \tag{15}$$

(2) The schematic diagram of the different DC modules is as follows:

As shown in Figure 5, this is the schematic representation of the various DC modules. Similarly, it traverses the feature map with a window size of 3×3, and the step length is 2. The differences between the corresponding pixels in the window are calculated. The pixel value corresponding to the point with the greatest difference is replaced by the pixel value of the entire window.

The preceding part is similar to the aforementioned DC module. The image is divided into a small window of 3×3. $A_i\left(i=1,2,3,\ldots,9\right)$ represents the values of nine points in the A window of the target image, and $B_i^{'}\left(i=1,2,3,\ldots,9\right)$ represents the values of nine points in the B' window of the contrast image. The distances $D_i^{'}\left(i=1,2,3,\ldots,9\right)$ between the corresponding points are then calculated, as shown in Formula (16).

$$D_i^{'}=\left|A_i-B_i^{'}\right|\left(i=1,2,3,\ldots,9\right) \tag{16}$$

The maximum value $D_{\max}^{'}$ in $D_i^{'}$ is obtained as shown in Formula (17).

$$D_{\max}^{'}=\max\left\{D_1^{'},D_2^{'},D_3^{'},\ldots,D_9^{'}\right\} \tag{17}$$

The corresponding pixel index value m' for $D_{\max}^{'}$ is shown in Formula (18).

$$m'=\Sigma f\left(i\right)\quad\left(i=1,2,3,\ldots,9\right) \tag{18}$$

The value of m' obtained in Formula (18) represents the index corresponding to the farthest point within the window between the

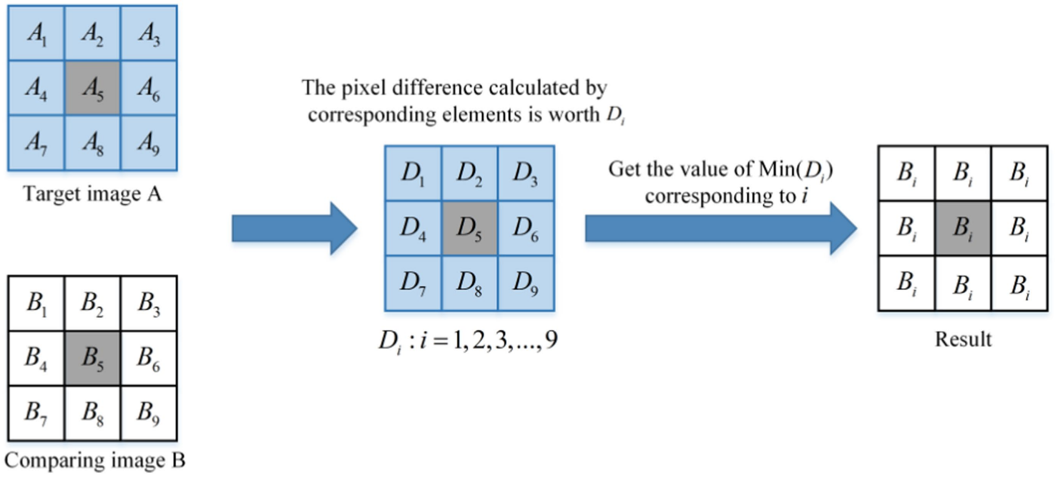


FIGURE 4
The principle of the similarity DC module.

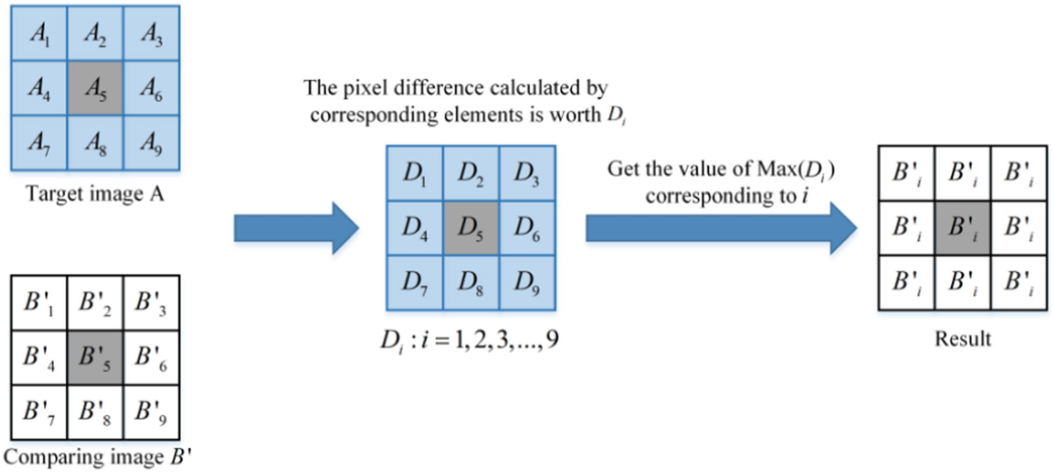


FIGURE 5
The principle of the difference DC module.

target image A and the contrast image B'. Here, $f(i)$ is defined as shown in Formula (19).

$$f(i) = \begin{cases} i, & D_i^{'} = D_{\max}^{'} \\ 0, & D_i^{'} \neq D_{\max}^{'} \end{cases} \tag{19}$$

Then, the values of all points in the window of contrast image B' is replaced with the value $B_{m'}$ corresponding to the point m', as shown in Formula (20).

$$B_i^{'} = B_{m'}, \quad (i = 1,2,3,\ldots,9) \tag{20}$$

After the sample pairs are processed, the corresponding similarity values are calculated, and the average value of the two is output as the final similarity coefficient.

For the similarity DC module, all eigenvalues in the window are replaced by the eigenvalues with the smallest distance between features in the feature map A. Similarly, for the differential module DC, all eigenvalues in the window are replaced by the eigenvalues with the widest distance between features in the feature map A. The average value of the corresponding elements in the two obtained features is then calculated to obtain the final feature A'.

2.2.4 Loss function

We train the model using a training set divided into batches. Each batch contains images $P \times K$, where P is the number of categories, and each category contains K images. First, three images are selected from the batches to feed the model, and a represents the current data, P is the image of the same category as a, and n is the image of a different type. Assuming that the sample is x and the total number of examples in the training set is N, the loss function *TriHard loss* is formulated in Formula (21).

$$\begin{gathered} TriHard\ loss = \frac{1}{N}\left\{\sum_{i}^{N}\left[d(a,p)_{\max} - d(a,n)_{\min} + \alpha\right]_{+}\right\} \\ d(a,p) = f\left(x_i^a\right) - f\left(x_i^p\right)_2^2 \\ d(a,n) = f\left(x_i^a\right) - f\left(x_i^p\right)_2^2 \end{gathered} \tag{21}$$

where $f(x)$ represents the mapping function of the model; max represents the maximum value; 'min' represents the minimum value; α represents the distance interval; the loss function makes the difference between $d(a,p)$ and $d(a,n)$ better than α.

3 Similarity metric

The core of vehicle re-identification tasks is to find and sort vehicle images. The ideal vehicle re-identification network model can make the distance metric between images of the exact vehicle smaller and more significant. In this study, the vehicle re-identification distance metric model is trained by vehicle models. As shown in Figure 6, in this study, according to the principle of metric learning based on the triple loss function, the distance metric between the anchor and the positive sample point becomes smaller and that between the anchor and the negative sample point becomes larger by training. This way, the recognition performance of re-identification of a vehicle with a positive sample is realized. The overall flow chart of this algorithm is shown in Figure 7.

4 Experimental results and analysis

4.1 Image dataset

The image data in this article comes from the VeRi776 datasets (Liu et al., 2016a) and VeRi-Wild datasets (Lou et al., 2019b). The

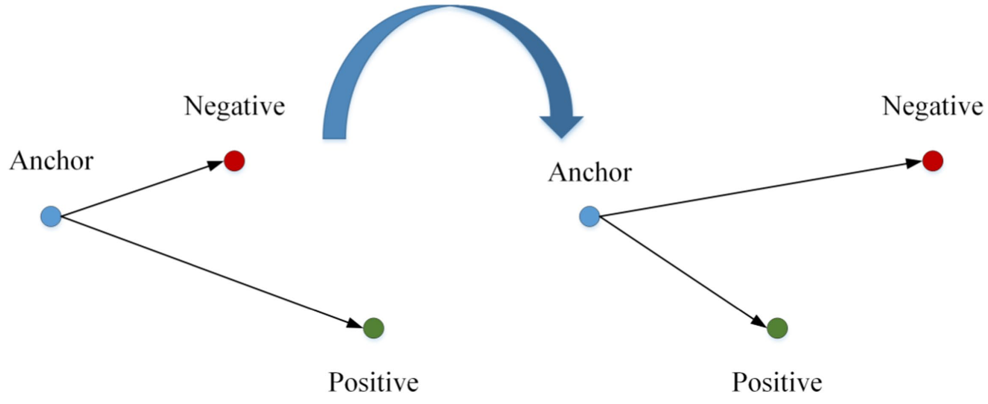


FIGURE 6
Metric learning method of triad loss function.

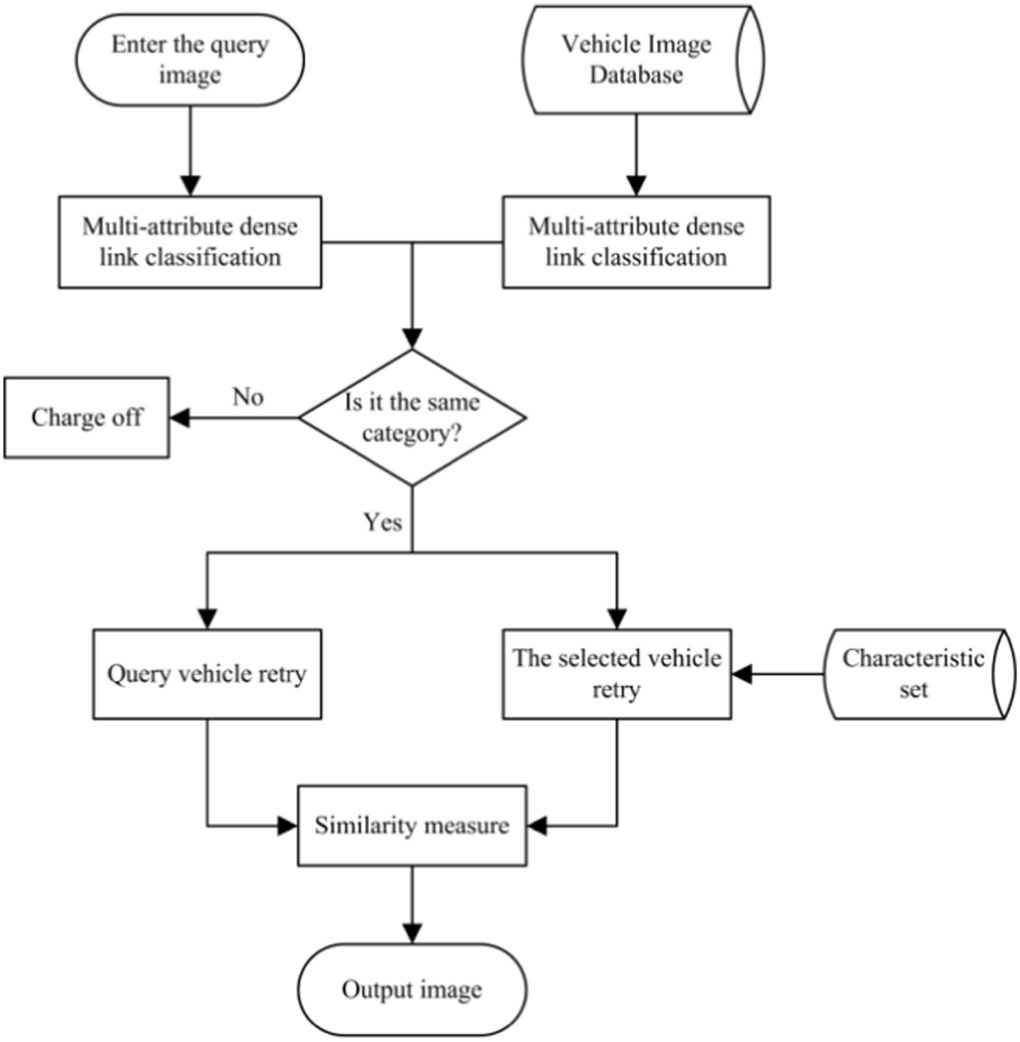


FIGURE 7
The overall flow chart of this algorithm.

VeRi776 dataset contains 50,000 images of 776 vehicles captured by 20 cameras without restrictions on traffic. Images of each vehicle are captured by 2 to 18 cameras with different viewing angles, illuminations, occlusions, and resolutions. Each vehicle image is tagged with vehicle ID and vehicle type information, with vehicle category information divided into nine categories. The dataset includes both a training set and a test set. The training set contains 37,782 images of 576 vehicles; the test set contains 13,257 images of 200 vehicles. To evaluate the results, the test dataset is further divided into a vehicle image library and test vehicle images. The vehicle image library contains 11,579 images of 200 vehicles, and the test vehicle image contains 1,678 images of 200 vehicles.

The imaging background and environmental variations of the VeRi-Wild dataset are more complex, and more camera models are used. The vehicle images in the dataset are captured by a 174-camera surveillance system covering more than 200 square kilometers. The total acquisition cost is 1 month. In total, more than 400,000 images of 40,000 vehicles were acquired (an average of 10 images per vehicle). In addition, the image angles included for the same vehicle vary widely. The dataset only annotates the vehicle ID without other information. This is the first dataset for vehicle re-identification without constraints. The sample images from the VeRi776 and VeRi-Wild datasets are shown in Figure 8.

4.2 Implementation details

We utilized the PyTorch framework to develop the network. The platform for training and testing is the Ubuntu 18.04 system, with a GTX 1070 Ti graphics card and 10GB of video memory. The hardware configuration for the experiments is presented in Table 1. This article uses the Adam optimizer with a learning rate of 0.001 and a momentum of 0.9. Since the neural network is very unstable at the beginning of training, a corresponding training strategy, cosine annealing learning, is added to reduce the risk of overfitting so that the model has strong robustness and good convergence to occlusion. In the cosine annealing strategy, the learning rate is reduced in the form of a cosine function, which ensures a smoother learning rate reduction and prevents the model from failing to converge because the learning rate is dropping too fast. The minimum learning rate is 0.00001.

The batch size is set to 32. We trained the network for 100 epochs. The experimental computer hardware configuration is shown in Table 1.

4.3 Experimental metric standard

In this experiment, $rank-n$, CMC curves, and mAP were used as experimental indexes to measure the model effect.

(1) *Accuracy*:

Accuracy is shown in Formula (22).

$$Accuracy = \frac{p}{N} \tag{22}$$

where p is the number of correctly identified samples, and N is the total number of samples.

(2) Classification rate v:

Using classification rate v to measure the speed of classification, the formula is shown in Formula (23).

$$v = n / s \tag{23}$$

where n is the number of classified samples, and s is the time needed to classify these pieces.

(3) $rank-n$ and CMC curve:

The result of vehicle re-identification is output as n images with the highest similarity between the test set and the query image. $rank-n$ represents the probability that the output of the first image,

after model determination, contains the correct image. For example, $rank-1$ represents the probability that the image with the highest similarity output, after model determination, is the correct image. $rank-5$ represents the probability that the first five image outputs, after model determination, contain the correct image.

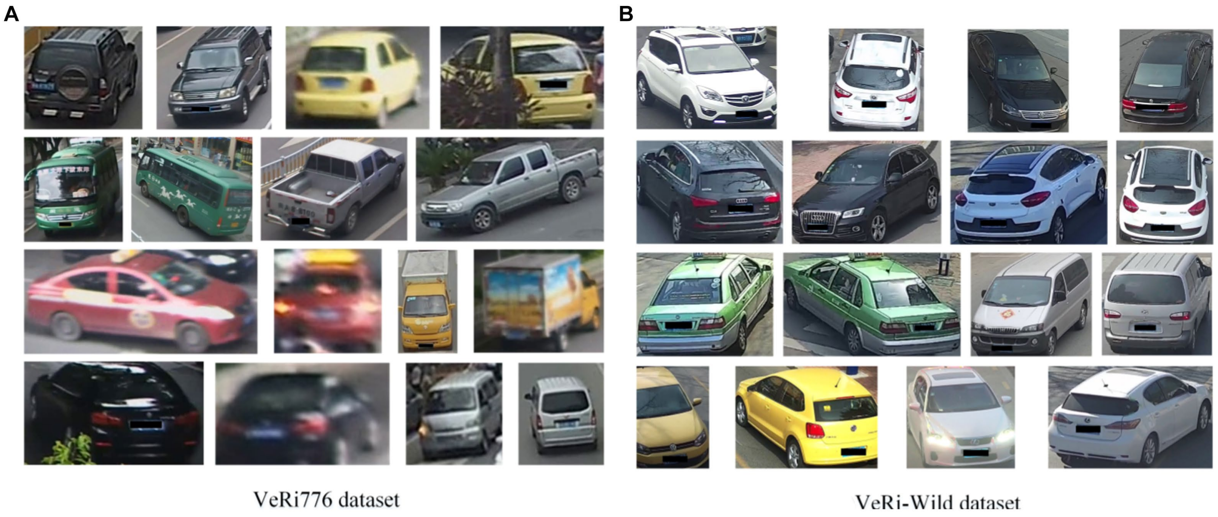


FIGURE 8
Some example images from both VeRi776 and VeRi-Wild datasets. **(A)** Example images from the VeRi776 dataset, **(B)** Example images from the VeRi-Wild dataset.

The CMC curve takes n value of $rank-n$ as abscissa, and the corresponding probability that the correct images are included, which is denoted by 8.

(4) mAP (Average precision rate):

The problem of vehicle re-identification is considered a two-class problem. The actual category of the requested image is considered as a positive category, and the false category is considered as a negative category. The identification results of the network are also divided into positive and negative categories.

$precision$ is calculated as shown in Formula (24).

$$precision=\frac{TP}{TP+FP} \tag{24}$$

The $precision$ represents the proportion of samples identified as positive classes in which the actual type is positive.

AP (Average Precision) represents average precision.

$$AP=\frac{1}{n}\sum_{i=1}^{n}p_i \tag{25}$$

In Formula (25), n denotes the number of right images returned, p_i denotes the corresponding precision of the first correct image.

When there are multiple re-identification objects, we average multiple AP values to mAP;

$$mAP=\frac{1}{N}\sum_{i=1}^{N}(AP)_i \tag{26}$$

In Formula (26), N denotes the number of re-identified objects. $(AP)_i$ means the AP value of the i re-identification object.

TABLE 1 Hardware equipment of practical environment.

Laboratory equipment	Experimental configuration
System	Ubantu18.04
Deep learning framework	Pytorch
Programming language	Python
Compiler	PyCharm
Running memory	32G
CPU	$Inter^{R}Core^{TM}i7-8750H,2.20GHz$
GPU	GTX1070Ti

4.4 Ablation experiment

In this section, we investigate the effectiveness of critical components in the mixed sample contrastive learning framework by conducting ablation studies on two different datasets. We introduced HSV features, type features, and the DC module into the network separately. Our proposed methodology aimed to enhance the differentiation between vehicles based on color and type, prompting the model to distinguish vehicles. Additionally, the DC module was utilized to shorten feature distances within image classes and expand feature distances between image classes. The experimental results demonstrated the significant effectiveness of multi-attribute features and the DC module in the context of vehicle re-identification tasks. The accuracy of the multi-attribute features composed of HSV color features and type features, along with the DC module, is presented in Table 2.

First, we incorporated the extraction of HSV color features for vehicle recognition into the network. Color is considered as a pivotal attribute for vehicles, enhancing the effectiveness of vehicle re-identification tasks. HSV color features serve to diminish the influence of image brightness on vehicle color recognition while also filtering out high saturation image elements such as windows and backgrounds that could otherwise interfere with color feature identification. Leveraging these color attributes, our model achieved an accuracy of 52.43% on VeRi-776 and 63.51% and 59.47% on VeRi-Wild (3000) and VeRi-Wild (5000), respectively.

Subsequently, vehicle-type features were introduced into the network. Type features assist in distinguishing visually similar vehicles. By leveraging type attributes, our model achieved an accuracy of 44.67% on VeRi-776 and 55.47 and 53.92% on VeRi-Wild (3000) and VeRi-Wild (5000), respectively.

Finally, we incorporated the distance control (DC) module into the network to assess its impact on accuracy. In networks featuring both HSV color and type features, the inclusion of the DC module resulted in our model achieving accuracies of 60.61% and 58.49% for VeRi-776, 67.34% and 63.97% for VeRi-Wild (3000), and 62.73% and 61.35% for VeRi-Wild (5000). The results in the seventh row show that the combined application of HSV color features, type features, and the DC module yields the highest mAP.

4.5 Experimental results and analysis

As shown in Table 3, the accuracy of CNN, VGG16, ResNet50, dense network, HSV + CNN, HSV + VGG16, HSV + ResNet50, and HSV + dense network is 83.17%, 86.47%, 91.78%, 90.63%, 88.76%,

TABLE 2 Experimental results of the vehicle classification method.

Algorithm	VeRi776-mAP	VeRi-Wild (3000)-mAP	VeRi-Wild (5000)-mAP
Backbone	42.54	50.16	49.72
Backbone + HSV	52.43	63.51	59.47
Backbone + type	44.67	55.47	53.92
Backbone + HSV + DC	60.61	67.34	63.97
Backbone + type + DC	58.49	62.73	61.35
Final	68.83	71.39	68.42

92.84%, 95.06%, and 94.24%, respectively. The classification efficiency is $50n/s$, $45n/s$, $63n/s$, $102n/s$, $47n/s$, $40n/s$, $55n/s$, and $94n/s$, respectively. In comparison, the accuracy of our algorithm ranks second, but compared with the first algorithm, the difference is only 0.82%, and the classification efficiency of this algorithm is higher than $39n/s$, so our algorithm is better than HSV+ ResNet50. As for the classification rate, our algorithm, although second, is only $8n/s$lower than the first one (dense connection network), yet the accuracy is 3.61% higher. The optimal accuracy and classification rate in comparative experiments, along with the results of the proposed method in this paper, have been bolded in Table 3. Therefore, in overall consideration, the accuracy and classification efficiency of the proposed algorithm are relatively optimal.

TABLE 3 Experimental results of the vehicle classification method.

Algorithm	Accuracy (%)	Classification efficiency(n/s)
CNN	83.17	50
VGG16	86.47	45
ResNet50	91.78	63
Densely connected network	90.63	**102**
HSV + CNN	88.76	47
HSV + VGG16	92.84	40
HSV + ResNet50	**95.06**	55
HSV + Densely connected network	**94.24**	**94**

Table 4 shows the comparison between the proposed algorithm model and the mainstream re-identification network on the VeRi776 dataset. Table 5 shows the comparison between the proposed algorithm and the mainstream algorithm on the VeRi-Wild dataset, using measures $rank-n$ and mAP.

From Table 4, it can be observed that, mAP, $rank-1$, and $rank-n$ of this algorithm achieve 68.83, 92.94, and 96.88%, respectively, and each index is the best. As for the second index, the mAP index is higher than the second by 0.18%, $rank-1$is higher by 2.84%, $rank-5$ is higher by 0.15%. As we can observe, this algorithm has the best performance among the above algorithms using VeRi776 dataset as the benchmark. As shown in Figure 9, the probability of classifying the first image output as the correct image is the highest compared with the other images. The advantage of this algorithm is that the hit rate of the model used in this study is relatively high compared with other algorithms.

Table 5 shows the experimental results of six different algorithms on the VeRi-Wild (3000) dataset and the VeRi-Wild (5000) dataset. It is obvious that due to the complex background and the angle of the vehicle images in the VeRi-Wild dataset, the overall performance of the index is lower than that of the VeRi776 dataset.

On the VeRi-Wild (3000) dataset, compared with the second-best CTCAL, the proposed algorithm outperforms by 1.04% in mAP, 2.73% in $rank-1$, and 1.76% in $rank-5$. In comparison to VAAG, which also utilizes multiple vehicle attributes for vehicle re-identification, the proposed algorithm demonstrates superiority by 2.14% in mAP, 3.10% in $rank-1$, and 0.92% in $rank-5$.

On the VeRi-Wild (5000) dataset, the proposed algorithm outperforms the second-best CTCAL by 2.69% in the mAP metric, 2.84% in the rank–1 metric, and 2.18% in the rank–5 metric.

TABLE 4 Experimental comparison of the VeRi776 dataset.

Models	*mAP* (%)	*rank–1* (%)	*rank–5* (%)
LOMO (Liao et al., 2015)	9.64	25.33	46.48
DGD (Xiao et al., 2016)	17.92	50.70	67.52
GoogLeNet (Yang et al., 2015)	17.81	52.12	66.79
FACT (Liu et al., 2016b)	18.73	51.85	67.16
Siamese Visual (Shen et al., 2017)	29.48	41.12	60.31
PAMAL (Tumrani et al., 2020)	45.06	-	-
MSCL (Yuefeng et al., 2022)	45.90	81.20	-
OIFE (Wang et al., 2017)	48.00	65.92	87.66
VAMI (Zhu et al., 2017)	50.13	77.03	90.82
QD-DFL (Zhu et al., 2020)	51.83	88.50	94.46
VRSDnet (Zhu et al., 2019)	53.45	83.49	92.55
FDA-Net (Lou et al., 2019a)	53.46	84.27	92.43
MV-GAN (Zhang et al., 2021)	61.16	91.06	95.77
VAAG (Tumrani et al., 2023)	63.01	92.20	96.64
VPEN (Meng et al., 2020)	67.98	90.36	94.84
VehicleNet (Zheng et al., 2021)	67.48	90.58	95.47
UFC (Wang et al., 2021)	68.24	91.84	96.73
CTCAL (Yu et al., 2021)	68.65	90.46	95.97
Ours	68.83	92.94	96.88

TABLE 5 Experimental comparison of the VeRi-Wild dataset.

Algorithm	VeRi-Wild (3000)			VeRi-Wild (5000)		
	mAP (%)	*rank–1* (%)	*rank–5* (%)	*mAP* (%)	*rank–1* (%)	*rank–5* (%)
GoogLeNet (Liao et al., 2015)	24.27	57.16	75.13	24.15	53.16	71.11
HDC (Yuan et al., 2017)	29.14	57.13	78.93	24.76	49.64	72.28
Unlabled GAN (Zhu et al., 2017)	29.86	58.06	79.60	24.71	51.58	74.42
FDA-Net (Lou et al., 2019a)	35.11	64.03	82.81	29.80	57.82	78.34
FDA-Net (Resnet50)	61.57	73.62	91.23	52.69	64.29	85.39
CTCAL (Yu et al., 2021)	70.35	83.64	92.63	65.73	80.31	90.75
Ours	71.39	86.37	94.39	68.42	83.15	92.93

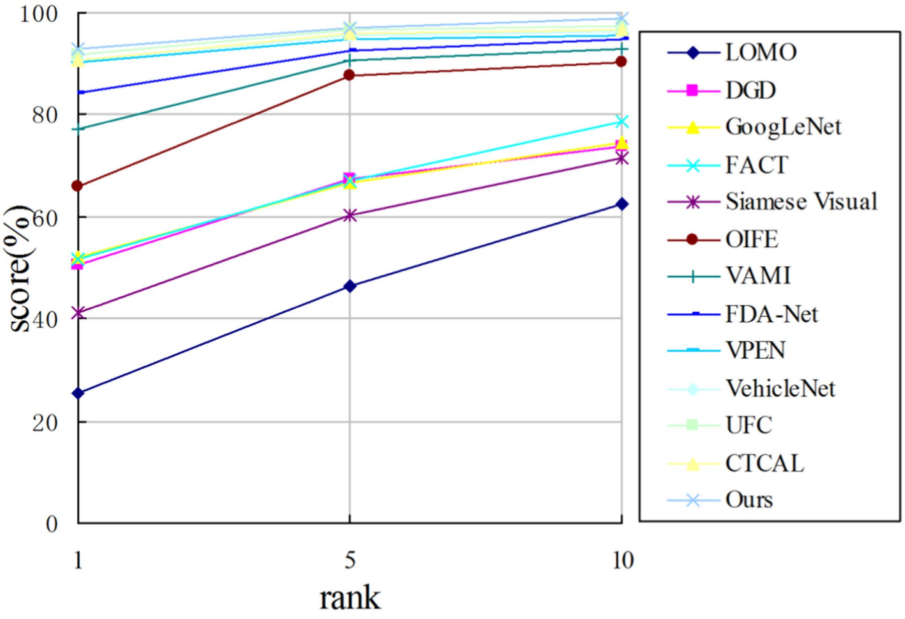


FIGURE 9
CMC curve on the VeRi776 dataset.

As shown in Figure 10, the algorithm in this study outperforms other algorithms in the metric*rank* − 1. With the increase of *n* value in the metric *rank* − *n*, the metric decreases gradually. Together with the experimental data in Table 4, it is proved that the algorithm in this study can distinguish the images with high similarity in the output images (the images in the foreground of the results) and improve the similarity between the classes.

The query images and ranking lists obtained by the final model on the VeRi776 dataset and VeRi-Wild are visually presented in Figures 11, 12. It can be observed that vehicles exhibit different appearances when subjected to varying perspectives, lighting conditions, and occlusions. Even during nighttime driving with illumination and reflection interference, the model can still recognize target images (Figure 12, last row). Overall, the experimental results indicate that the proposed method outperforms existing state-of-the-art multi-attribute-based vehicle re-identification methods.

5 Conclusion

In this study, we propose a vehicle re-identification method that integrates a multi-attribute dense connection network with a distance control (DC) module. This model introduces a multi-attribute dense connection mechanism based on HSV color attributes and category attributes in the feature extraction segment of the network, reducing computational complexity. Feature extraction is achieved through multi-dimensional feature-weighted fusion, enhancing both feature extraction and classification accuracy. Furthermore, a method controlling inter-category distances is introduced, employing a DC module as an image distance control module. This module comprises both similar and different DC modules. Based on the target image, features of input images are processed through both similarity and difference DC modules, and the resulting features are then merged into the subsequent network for similarity determination. This module

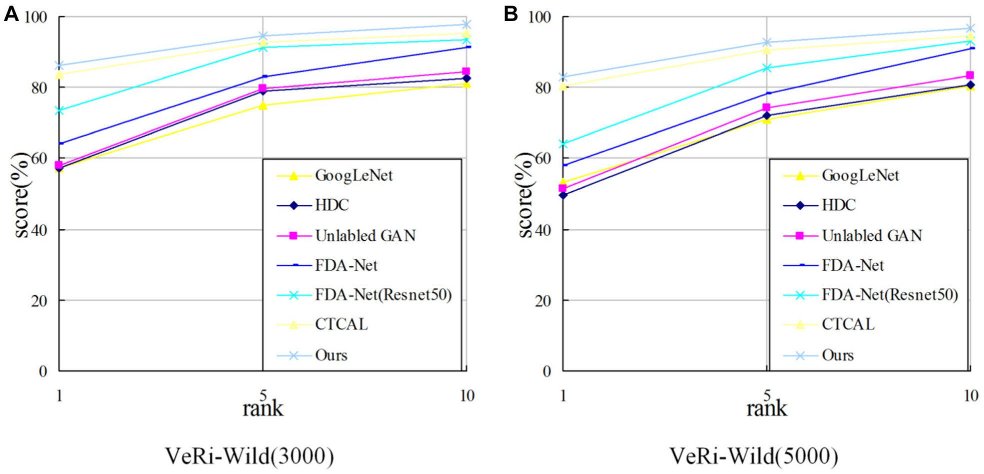


FIGURE 10
CMC comparison on the VeRi-Wild dataset. **(A)** CMC comparison on the VeRi-Wild (3000); **(B)** CMC comparison on the VeRi-Wild (5000).

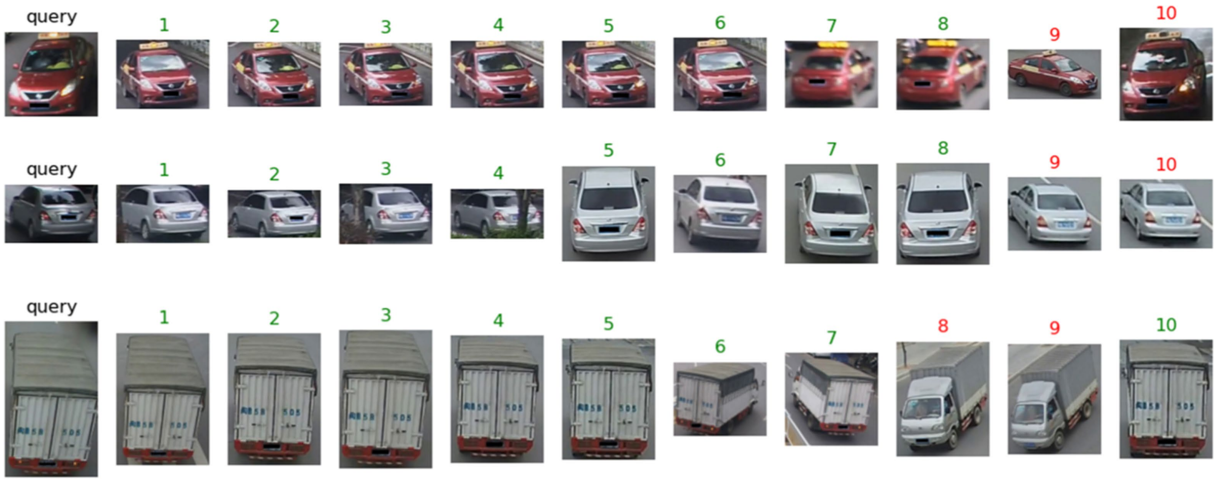


FIGURE 11
The proposed model results on the VeRi776 dataset. The green numbers and red numbers illustrate the correct and wrong matches.

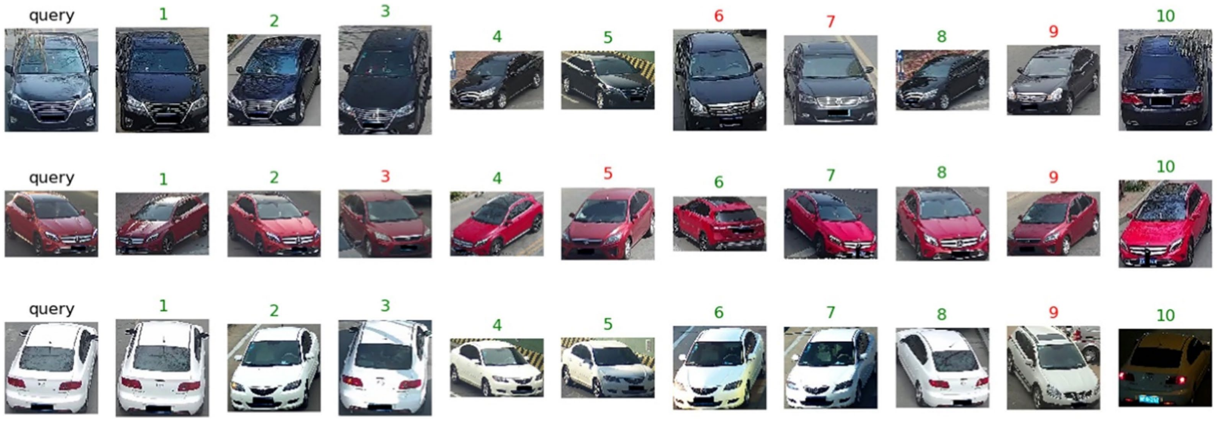


FIGURE 12
The proposed model results on the VeRi-Wild dataset. The green numbers and red numbers illustrate the correct and wrong matches.

effectively reduces feature distances within image categories while increasing distances between image categories, thereby elevating the accuracy of vehicle re-identification. Experiments for vehicle re-identification are conducted using the VeRi776 dataset, yielding precision and recall values of 68.83 and 92.94%, respectively, surpassing values obtained by other comparative algorithms. Further experiments using the VeRi-Wild (3000) and VeRi-Wild (5000) datasets for vehicle re-identification demonstrate precision values of 71.39% and 68.42% and recall values of 86.37% and 83.15%, respectively, outperforming other algorithms. Experimental results affirm the efficacy of the proposed method in enhancing the accuracy of vehicle re-identification.

Data availability statement

The data analyzed in this study is subject to the following licenses/ restrictions: the production personnel of this dataset ask for your information only to make sure the dataset is used for non-commercial purposes. They will not give it to any third party or publish it publicly anywhere. Requests to access these datasets should be directed to VeRi776 datasets, https://vehiclereid.github.io/VeRi/ and VeRi-Wild datasets, https://github.com/PKU-IMRE/VERI-Wild.

Author contributions

XS: Methodology, Software, Writing – review & editing. YC: Methodology, Software, Writing – original draft. YD: Methodology, Software, Writing – original draft. YW: Methodology, Software, Writing – original draft. JZ: Methodology, Software, Writing – original draft. BS: Supervision, Writing – review & editing. LL: Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was Supported by the Scientific Research Startup Fund for Shenzhen High-Caliber Personnel of SZPT (6023330002K), General Higher Education Project of Guangdong Provincial Education Department (2023KCXTD077), Guangdong Provincial General University Innovation Team Project (2020KCXTD047), college start-up fund of ShenZhen PolyTechnic University (6022312031K).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abdulhai, B., and Tabib, S. M. (2003). Spatio-temporal inductance-pattern recognition for vehicle re-identification. *Transport Res Part C Emerg Technol* 11, 223–239. doi: 10.1016/S0968-090X(03)00024-X

Bashir, R. M. S., Shahzad, M., and Fraz, M. M. (2019). Vr-proud: vehicle re-identification using progressive unsupervised deep architecture. *Pattern Recogn* 90, 52–65. doi: 10.1016/j.patcog.2019.01.008

Coifman, B. (1998). Vehicle re-identification and travel time measurement in real-time on freeways using existing loop detector infrastructure. *Transp Res Rec* 1643, 181–191. doi: 10.3141/1643-22

Guo, H., Zhu, K., Tang, M., and Wang, J. (2019). Two-level attention network with multi-grain ranking loss for vehicle re-identification [J]. *IEEE Trans. Image Process.* 28, 4328–4338. doi: 10.1109/TIP.2019.2910408

Hou, J., Zeng, H., Zhu, J., Hou, J., Chen, J., and Ma, K. K. (2019). Deep quadruplet appearance learning for vehicle re-identification. *IEEE Trans Veh Technol* 68, 8512–8522. doi: 10.1109/TVT.2019.2927353

Jin, Y., Li, C., Li, Y., Peng, P., and Giannopoulos, G. A. (2021). Model latent views with multi-center metric learning for vehicle re-identification. *IEEE Trans Intell Transp Syst* 22, 1919–1931. doi: 10.1109/TITS.2020.3042558

Liao, S., Hu, Y., Zhu, X., and Li, S. (2015). "Person re-identification by local maximal occurrence representation and metric learning", In: *Proceedings IEEE Conference Computing Vision and Pattern Recognition*. 2197–2206. arXiv [Preprint]. arXiv:1406.4216v2]

Liu, X., Liu, W., Ma, H., and Fu, H. (2016a). "Large-scale vehicle re-identification in urban surveillance videos", ICME, 1–6.

Liu, X., Liu, W., Mei, T., and Ma, H. (2016b). "A deep learning-based approach to progressive vehicle re-identification for urban surveillance", In: *European conference on computer vision*. Springer, Cham, 9906, 869–884.

Lou, Y., Bai, Y., Liu, J., Wang, S., and Duan, L. (2019a). Veri-wild: a large dataset and a new method for vehicle re-identification in the wild. *CVPR* 2019a, 3235–3243. doi: 10.1109/CVPR.2019.00335

Lou, Y., Bai, Y., Liu, J., Wang, S., and Duan, L. (2019b) Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3235-3243.

Meng, D., Liang, L., Liu, X., Li, Y., Yang, S., Zha, Z., et al. (2020). "Parsing-based view-aware embedding network for vehicle re-identification", arXiv [Preprint]. arXiv: 2004.05021.

Peng, J., Wang, H., Xu, F., and Fu, X. (2020). Cross domain knowledge learning with dual-branch adversarial network for vehicle re-identification. *Neurocomputing* 401, 133–144. doi: 10.1016/j.neucom.2020.02.112

Qian, J., Jiang, W., Luo, H., and Yu, H. (2020). Stripe-based and attribute-aware network: a two-branch deep model for vehicle re-identification. *J Phys E Sci Instr* 31:095401. doi: 10.1088/1361-6501/ab8b81

Ratnesh, K., Edwin, W., Farzin, A., and Parthasarathy, S. (2020). A strong and efficient baseline for vehicle re-identification using deep triplet embedding. *J Artif Intell Soft Comput Res* 10, 27–45. doi: 10.2478/jaiscr-2020-0003

Shen, Y., Xiao, T., Li, H., Yi, S., and Wang, X. (2017). "Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals", In: *Proceedings of the IEEE International Conference on Computer Vision*. 1900-1909.

Teng, S., Zhang, S., Huang, Q., and Sebe, N. (2020). Multi-view spatial attention embedding for vehicle re-identification. *IEEE Trans Circuits Syst Video Technol* 31, 816–827. doi: 10.1109/TCSVT.2020.2980283

Tumrani, S., Ali, W., Kumar, R., Khan, A. A., and Dharejo, F. A. (2023). View-aware attribute-guided network for vehicle re-identification. *Multimedia Syst* 29, 1853–1863. doi: 10.1007/s00530-023-01077-y

Tumrani, S., Deng, Z., Lin, H., and Shao, J. (2020). Partial attention and multi-attribute learning for vehicle re-identifcation. *Pattern Recogn. Lett.* 138, 290–297. doi: 10.1016/j.patrec.2020.07.034

Wang, P., Ding, C., Tan, W., Gong, M., Jia, K., and Tao, D. (2021) "Uncertainty-aware clustering for unsupervised domain adaptive object re-identification", arXiv [Preprint]. arXiv: 2108.09682.

Wang, Z., Tang, L., Liu, X., Yao, Z., Yi, S., Shao, J., et al. (2017) "Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification", Proceedings of the IEEE International Conference on Computer Vision, 379–387.

Xiao, T., Li, H., Ouyang, W., and Wang, X. (2016). "Learning deep feature representations with domain guided dropout for person re-identification", In: *Proceedings IEEE Conference Computing Vision and Pattern Recognition*. 1249–1258.

Yang, L., Luo, P., Loy, C., and Tang, X. (2015). "A large-scale car dataset for fine-grained categorization and verification", In: *Proceedings IEEE Conference Computing Vision and Pattern Recognition*. 3973–3981. arXiv [Preprint]

Yu, J., Kim, J., Kim, M., and Oh, K (2021) "Camera-Tracklet-aware contrastive learning for unsupervised vehicle re-identification", arXiv [Preprint] arXiv: 2109.06401

Yuan, Y., Yang, K., and Zhang, C. (2017). "Hard-aware deeply cascaded embedding", Proceedings of the IEEE International Conference on Computer Vision, 814–823.

Yuefeng, W., Ying, W., Ruipeng, M., and Lin, W. (2022). Unsupervised vehicle re-identifcation based on mixed sample contrastive learning. *Signal Image Video Process* 16, 2083–2091. doi: 10.1007/s11760-022-02170-x

Zhang, J., Chen, J., Cao, J., Liu, R., Bian, L., and Chen, S. (2022). Dual attention granularity network for vehicle re-identification. *Neural Comput. Applic.* 34, 2953–2964. doi: 10.1007/s00521-021-06559-6

Zhang, F., Ma, Y., Yuan, G., Zhang, H., and Ren, J. (2021). Multiview image generation for vehicle reidentifcation. *Appl Intell* 51, 5665–5682. doi: 10.1007/s10489-020-02171-8

Zhang, X., Zhang, R., Cao, J., Gong, D., You, M., and Shen, C. (2019). Part-guided attention learning for vehicle re-identification]. arXiv [Preprint], arXiv: 1909.06023.

Zheng, Z., Ruan, T., Wei, Y., Yang, Y., and Mei, T. (2021). VehicleNet: learning robust visual representation for vehicle re-identification. *IEEE Trans. Multimed.* 23, 2683–2693. doi: 10.1109/TMM.2020.3014488

Zhu, J., Du, Y., Hu, Y., Zheng, L., and Cai, C. (2019). Vrsdnet: vehicle reidentifcation with a shortly and densely connected convolutional neural network. *Multimed. Tools Appl.* 78, 29043–29057. doi: 10.1007/s11042-018-6270-4

Zhu, J. Y., Park, T., Isola, P., and Efros, A.. (2017). "Unpaired image-to-image translation using cycle-consistent adversarial networks", In: *Proceedings of the IEEE International Conference on Computer Vision*, 2223–2232.

Zhu, J., Zeng, H., du, Y., Lei, Z., Zheng, L., and Cai, C. (2018). Joint feature and similarity deep learning for vehicle re-identification. *IEEE Access* 6, 43724–43731. doi: 10.1109/ACCESS.2018.2862382

Zhu, J., Zeng, H., Huang, J., Liao, S., Lei, Z., Cai, C., et al. (2020). Vehicle re-identifcation using quadruple directional deep learning features. *IEEE Trans Intell Transp Syst* 21, 410–420. doi: 10.1109/TITS.2019.2901312

frontiers | Frontiers in Neurorobotics

TYPE Original Research
PUBLISHED 16 January 2024
DOI 10.3389/fnbot.2024.1293992

OPEN ACCESS

EDITED BY
Long Jin,
Lanzhou University, China

REVIEWED BY
Zhenghua Huang,
Wuhan Institute of Technology, China
Tianwen Zhang,
University of Electronic Science and Technology of China, China

*CORRESPONDENCE
Chenke Yue
✉ yueck0928@nuaa.edu.cn
Chao Li
✉ 19b921033@stu.hit.edu.cn

RECEIVED 14 September 2023
ACCEPTED 03 January 2024
PUBLISHED 16 January 2024

CITATION
Li C, Yue C, Li H and Wang Z (2024)
Context-aware SAR image ship detection and recognition network.
Front. Neurorobot. 18:1293992.
doi: 10.3389/fnbot.2024.1293992

Context-aware SAR image ship detection and recognition network

Chao Li[1]*, Chenke Yue[2,3]*, Hanfu Li[1] and Zhile Wang[1]

[1]School of Astronautics, Harbin Institute of Technology, Harbin, Heilongjiang, China, [2]Key Laboratory of Space Photoelectric Detection and Perception (Nanjing University of Aeronautics and Astronautics), Ministry of Industry and Information Technology, Nanjing, Jiangsu, China, [3]Nanjing University of Aeronautics and Astronautics, College of Astronautics, Nanjing, Jiangsu, China

With the development of deep learning, synthetic aperture radar (SAR) ship detection and recognition based on deep learning have gained widespread application and advancement. However, there are still challenging issues, manifesting in two primary facets: firstly, the imaging mechanism of SAR results in significant noise interference, making it difficult to separate background noise from ship target features in complex backgrounds such as ports and urban areas; secondly, the heterogeneous scales of ship target features result in the susceptibility of smaller targets to information loss, rendering them elusive to detection. In this article, we propose a context-aware one-stage ship detection network that exhibits heightened sensitivity to scale variations and robust resistance to noise interference. Then we introduce a Local feature refinement module (LFRM), which utilizes multiple receptive fields of different sizes to extract local multi-scale information, followed by a two-branch channel-wise attention approach to obtain local cross-channel interactions. To minimize the effect of a complex background on the target, we design the global context aggregation module (GCAM) to enhance the feature representation of the target and suppress the interference of noise by acquiring long-range dependencies. Finally, we validate the effectiveness of our method on three publicly available SAR ship detection datasets, SAR-Ship-Dataset, high-resolution SAR images dataset (HRSID), and SAR ship detection dataset (SSDD). The experimental results show that our method is more competitive, with AP50s of 96.3, 93.3, and 96.2% on the three publicly available datasets, respectively.

KEYWORDS

ship detection, synthetic aperture radar (SAR), channel-wise attention, context-aware, aggregation

1 Introduction

SAR is an active microwave imaging sensor, which can obtain high-resolution radar images under low visibility weather conditions, and it is widely used in the field of ship monitoring (Yang et al., 2018), geological exploration (Ghosh et al., 2021), and climate forecasting (Mateus et al., 2012). Distinguished from other remote sensing modalities, SAR stands out due to its ability to operate day and night, under all weather conditions, and its high resolution. So it makes SAR a crucial tool for object detection and marine monitoring. Recently, scholars have shown significant interest in utilizing SAR for ship detection in ports and on the open sea, and its applications have proven vital in both military and civilian domains.

In the past decades, a series of traditional SAR ship detection methods have emerged as the research related to SAR imaging technology and surface ship detection has been continuously and vigorously developed. The most representative types of traditional methods, such as the global threshold-based method that determines a global threshold through statistical decision-making and then searches for bright spot targets in the whole SAR image (Eldhuset, 1996), adaptive threshold methods that utilize the statistical distribution of sea clutter to determine an adaptive threshold with a constant false alarm probability (Rohling, 1983) and generalized likelihood ratio methods that take into account the distributional properties of both the background clutter and the ship's target (Iervolino and Guida, 2017). However, these traditional methods are based on interpretable theoretical justifications and well-established a priori knowledge to analyze ship features in SAR images, relying on manual feature extraction. When facing complex backgrounds and SAR images with a small proportion of target pixel values, the use of manually predefined features proves challenging in extracting effective target information and eliminating background noise interference. This results in a high false negative rate in target detection, preventing the accurate identification of ship targets. With the development of convolutional neural network (CNN) and the emergence of extensive SAR image ship detection datasets, such as SAR-Ship-Dataset (Wang et al., 2019), HRSID (Wei et al., 2020), and SSDD (Li et al., 2017), which has led to the rapid development of remote sensing image-based SAR target detection techniques for ships, especially in the feature extraction of targets.

Initially, driven by a substantial quantity of publicly SAR ship datasets, several deep learning-based multi-target detectors were directly used in SAR ship detection tasks. Such as two-stage detectors, region extraction-based convolutional neural networks (RCNN; Girshick et al., 2014), FastRCNN (Girshick, 2015) and the FasterRCNN, which is representative (Ren et al., 2015). Another example is single-stage detectors such as RetinaNet (Lin et al., 2017a), SSD (Liu et al., 2016), CenterNet (Zhou et al., 2019), and YOLO series (Redmon et al., 2016; Redmon and Farhadi, 2017, 2018). The above algorithms can automatically mine the effective features of the target and no longer rely on manual extraction, but they are ineffective, those who were initially designed for use as a general-purpose object detector in visible light. Subsequently, many scholars began to consider the design of deep networks for the task of ship target detection in SAR images. For example, Ma et al. (2022) proposed a ship target detection method based on attention mechanism and key point estimation. The method uses residual link and hierarchical features to extract multi-scale targets, then uses an attention mechanism to focus on target features and detect key points to solve the dense arrangement problem. As for multi-scale problem, Zhang et al. (2022) expanded the scope of image perception region by acquiring multiple scale slices with different region sizes. In addition, they addressed the issue of false positives by calculating the distinctiveness between targets and background, and by employing a multi-ensemble reasoning mechanism to merge confidence scores from multiple bounding boxes, which enhanced the extraction of target features.

Quad-FPN (Zhang et al., 2021a) sequentially concatenated four distinct feature pyramid network (FPN; Lin et al., 2017b), progressively enhancing detection performance. Yang et al. (2021) designed the Coordinate Attention Module (CoAM), embedding positional information into channels, thereby enhancing sensitivity to spatial details and strengthening the localization of ship targets. Then, they designed the receptive field increased module (RFIM), which employs multiple parallel convolutions to construct a spatial pyramid structure, to acquire multi-scale target information.

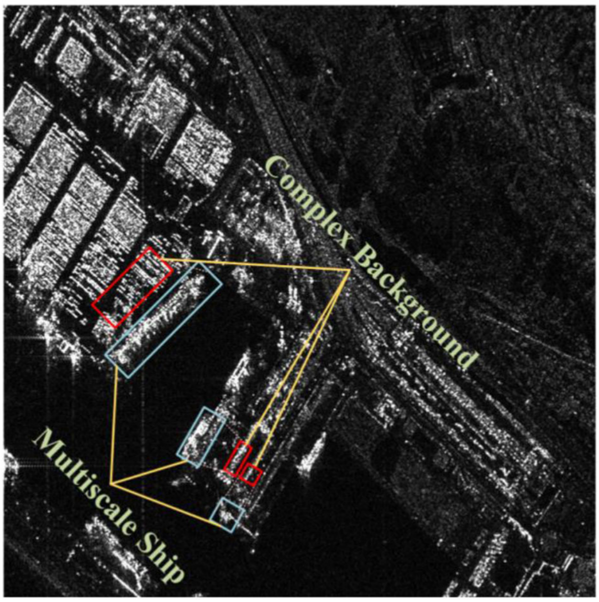


FIGURE 1
Examples of SAR images with complex backgrounds and different scales in the SAR ship dataset. The blue boxed lines show ships of different scale sizes, and the red boxed lines show the complex background noise interference that their ships may be subjected to around them.

However, in practical applications, numerous challenging issues exist, as illustrated in Figure 1. On one hand, due to the coherent imaging principles in SAR images, adjacent pixel values undergo random variations, leading to speckle noise in the image. In scenarios such as coastal ports, islands, and regions with sea clutter, SAR ship images may struggle to extract valid information, resulting in instances of both missed detections and false positives. On the other hand, the multiscaling problem poses another challenge. The varying resolutions and morphological sizes of ship targets necessitate higher demands for multiscale feature extraction from the network model, given that the pixel range occupied by ship targets can vary from a few to several hundred.

Firstly, to address the issue of significant scale variations in ship targets, we designed a LFRM, which improves upon atrous spatial pyramid pooling (ASPP; Chen et al., 2017). Apart from the first layer, a residual link is employed for each atrous convolution layer to receive and fuse the output from the previous layer, concatenating it with the current layer's output. This effectively integrates information from different scales. Finally, by combining a dual-branch channel attention mechanism using global average pooling (GAP) and global max pooling (GMP), we achieve local cross-channel interactions. The overall network architecture of our proposed method employs a multi-level design with multiple detection heads to detect targets of different sizes, making it more suitable for multiscale targets.

Secondly, to mitigate the impact of noise from a complex background on the target, we introduce the GCAM, which expands the network's sensory domain by adaptively weighting features in different spaces. It leverages estimation-based long-range dependencies to obtain global semantic features, concentrating on the target's intrinsic characteristics to weaken background noise interference. Finally, we sequentially link and embed these two modules into the Feature Pyramid Network (FPN; Lin et al., 2017b) structure with a backbone network, enabling multi-level, wide-angle perception of context. The main contributions of this paper are as follows:

- We propose a context-aware SAR image ship detection and recognition network (CANet) that effectively detects multiscale targets through both bottom-up and top-down pathways, equipped with multiple detection heads.
- A Local Feature Refinement Module (LFRM) is designed to acquire target features of varying receptive field sizes, enabling local cross-channel interactions to enhance the model's performance.
- We introduce a GCAM to capture long-range dependencies, perceive global context, strengthen target representation, and suppress noise.
- To validate the effectiveness of our approach, extensive experiments were conducted on several authoritative SAR ship detection datasets, including SAR-Ship-Dataset (Wang et al., 2019), HRSID (Wei et al., 2020), and SSDD (Li et al., 2017). Our method demonstrated outstanding performance with detection accuracies reaching 96.3, 93.3, and 96.2%, respectively.

2 Related work

SAR image ship target detection methods are mainly categorized into traditional methods and deep learning-based methods. The former defines ship target features manually, and then search for feature-matched ship targets in SAR images based on the predefined features, which can be categorized into three main groups: based on transform domain (Schwegmann et al., 2016), threshold-based algorithms (Renga et al., 2018) and statistical feature distribution algorithms (Wang et al., 2013). Within, the most representative one is the constant false alarm-based (CFAR-based) method. It is based on the statistical model of sea clutter, which is affected by the ocean area, the wind field conditions of the ocean, and the radar backscattering intensity varies in different wind field regions, thus forming a more complex clutter edge environment at the junction of different regions. Therefore, it is challenging to establish an accurate statistical model for a wide range of complex sea clutter. In addition, clutter modeling often requires complex mathematical theory support and time-consuming manual involvement, which also reduces the flexibility of the model and makes it difficult to effectively detect ship targets.

In recent years, convolutional neural networks (CNNS) have made great achievements in the field of natural image object detection, and their detection performance has been significantly improved compared with traditional methods. At present, natural image object detection methods based on deep learning are mainly divided into two categories: single-level object detectors and two-level object detectors. Girshick et al. (2014) proposed the first two-stage target detection model, R-CNN, which employs a traditional selective search algorithm to generate about 2,000 candidate frames, which are then fed into the CNN to extract features and categorize the candidate frames, and finally obtain the detection results. Subsequently, inspired by SPPNet (He et al., 2015), Fast R-CNN (Girshick, 2015) was proposed to solve the problem of slow detection speed of RCNN, which extracts the ROI features on the network feature map to avoid the repeated computation of features. It improved the detection speed. They used the Fully Connected (FC) layer instead of the original SVM classifier to further improve the classification performance. Ren et al. (2015), who proposed the faster FasterRCNN, designed the RPN network to replace the traditional candidate region generation algorithm selective search (Uijlings et al., 2013), which uses the convolutional network to extract the features and generate the position of the pre-selected frame. It reduces the time burden caused by the selective search algorithm and can almost reach the standard of real-time detection. More recently, faster R-CNN (Ren et al., 2015) is still the mainstream representative of two-stage detectors, and its mature design scheme has been widely used by numerous scholars.

As more demanding real-time target detection tasks are proposed, single-stage target detection is developing rapidly. As the pioneers of single-level target detectors, the YOLO series (Redmon et al., 2016; Redmon and Farhadi, 2017, 2018), by directly treating the object detection problem as the regression problem of the target region position and target category prediction, can output the positions and categories of target bounding boxes using only convolutional networks, meeting the requirement of real-time detection. Subsequently, YOLOv4 (Bochkovskiy et al., 2020) and YOLOv5 were proposed to achieve a new balance between the accuracy and speed of this series of algorithms, which were applied to more detection and recognition tasks. Another improvement of YOLO, TPH-YOLO (Zhu et al., 2021), to improve the detection accuracy of tiny targets, a tiny target detection head is added based on YOLOv5, and a total of four Prediction heads can mitigate the effects of large changes in the size of the target scale. Meanwhile, it replaces some convolutional blocks with transformer encoder ones to capture global information and sufficient background semantic information. SSD (Liu et al., 2016) and RetinaNet (Lin et al., 2017a) are two other common single-stage detectors. The former directly utilizes convolutional layers to extract detection results from different feature maps. It employs prior boxes with varying scales and aspect ratios to better match the shapes of targets, distinguishing it from YOLO, which uses fully connected layers for detection. While the latter proposes a new loss function that can be used as a more efficient alternative to previous methods for dealing with class imbalance. This class imbalance problem is solved by reshaping the standard cross-entropy loss to reduce the loss assigned to well-categorized examples.

With the blooming of deep learning in the field of images, CNN-based ship detection is increasingly subject to becoming popular. Dense Attention Pyramid Network (DAPN; Cui et al., 2019) embedded a convolutional block attention module (CBAM) into each level of the pyramid structure from the bottom up to enrich the semantic information on different level scale features and amplify the significance of features. CBAM is used to fuse the features at all levels, and the adaptive selection focuses on the scale features to further strengthen the detection and recognition of multi-scale targets. Also improved based on FPN (Lin et al., 2017b), Zhao et al. presented a novel network called attention receptive pyramid network (ARPN; Zhao et al., 2020), by fine-tuning the pyramid structure, to generate candidate boxes at different levels of the pyramid. Then, asymmetric convolution and atrous convolution are used to obtain convolution features in different directions to enhance the global context features of the local region. Then channel attention and space attention are combined to re-weight the extracted features, improving the significance of the target features and suppressing the interference of noise, and finally connect them to each layer of the pyramid laterally. Chaudhary et al. (2021) tried to directly apply YOLOv3 (Redmon and Farhadi, 2018) to ship detection and achieved some good results. Inspired by YOLO, Zhang and Zhang (2019) divided the original image into grid regions, and each grid was independently responsible for detecting the target in the region. Then, the image features are extracted through the backbone network for detection. In particular, backbone networks use separable convolution to reduce network burden.

PPA-Net (Tang et al., 2023) took into consideration that the designs of attention mechanisms such as CBAM are tailored for natural images, overlooking the impact of speckle noise in SAR images on attention weight generation. The target salience information is introduced into the attention mechanism to obtain the attention weight suitable for the SAR image. First, three pooled operations of different region sizes are constructed to obtain parallel multi-scale branches, and then activation functions are used to obtain the final channel attention weights. Meanwhile, considering the mutual exclusivity between semantic and location information and avoiding simple feature cascade operations, the authors use two self-attention weights to adaptively regulate the fusion feature ratio. To enhance the practical value of SAR ship detection applications, Zhang et al. (2019) constructed a lightweight SAR ship detection network based on the depthwise separable convolution neural network (DS-CNN). They replaced traditional convolutions with DS-CNN, significantly improving detection speed with fewer parameters, making it applicable for real-time detection tasks. Similarly, to improve detection speed, Lite-yolov5 (Xu et al., 2022a) designed a lightweight stride module and pruned the model to create a lightweight detector. To ensure detection accuracy, histogram and clustering methods were applied to enhance detection performance. Additionally, there are instance segmentation methods based on SAR ships, such as the attention interaction and scale enhancement network (MAI-SE-Net; Zhang and Zhang, 2022a). This method models long-range dependencies to enhance global perception and uses feature recombination to generate high-resolution feature maps, improving the detection capabilities for small targets. Zhang and Zhang (2022b) employed a dense sampling strategy, fusing features extracted by FPN at each layer and adding contextual information to the region of interest (ROI) to enhance information gain.

To address the issue of multiscale object detection, HyperLi-Net (Zhang et al., 2020a) utilized five improved internal modules to enhance the accuracy of multiscale object detection. These modules include multiple receptive fields, dilated convolution, attention mechanisms, and a feature pyramid to extract multiscale contextual information. Xu et al. (2022b) utilized the polarimetric characteristics of SAR to enhance feature expression and fused multiscale polarimetric features to obtain scale information. Zhang and Zhang (2020) proposed a lightweight one-stage SAR ship detection method, ShipDeNet-20. Because it uses depth-separable convolution with fewer layers and parameters instead of traditional convolution, its detection speed and model size are superior to other detection methods. Meanwhile, to ensure that the detection accuracy is not lost, features of different depths are fused to enhance the contextual semantics of features, and feature maps of the same size are superimposed to improve the expression ability of features, to improve the detection accuracy. Zhu et al. (2022) used the gradient density parameter g to construct the loss function of the network in order to solve the sparse problem of ship targets unbalanced with positive and negative ship samples. To prevent positive samples from having a decisive influence on the global gradient, the weight of the gradient proportion of multiple samples is neutralized. The author also studies the effect of the imbalance of feature levels on multi-scale ship detection. In order to ensure that semantic information is not lost during multi-layer transmission, the method of horizontal link integration of multilevel features is adopted to accelerate the flow of information so that the detailed features and semantic features can achieve balance, avoiding the semantic information and detailed features caused by the loss of other resolutions only by focusing on adjacent resolution information.

To mitigate the impact of background noise on the target, the Balance Scene Learning Mechanism (BSLM; Zhang et al., 2020b) employs a generative adversarial network (GAN) to extract complex scene information from SAR. This is followed by a clustering method to differentiate between nearshore and offshore backgrounds, thus enhancing the background. Similar balancing strategies are employed in various methods (Zhang et al., 2020c, 2021b). Additionally, some approaches utilize pixel-level processing to reduce background noise. Sun et al. (2023) used superpixels to reduce the impact of noise on the target. Firstly, the image is segmented by pixel blocks of different sizes to obtain target features of different sizes and image understanding of different semantic levels. After that, the surrounding contrast feature region is dynamically selected by dividing the size of the superpixel so that the smaller superpixel can have a larger contrast region while the larger superpixel can choose the features around itself for comparison. Finally, the superpixel features at different levels are fused for detection. Previous studies focused on extracting the features of ship targets in the spatial domain, but Li et al. (2021) believed that the spatial features of ship targets could not meet the requirements of high-precision detection, so they used the frequency domain to make up for the shortcomings in the spatial domain. Like most methods, the multi-scale spatial information of the ship target space domain is obtained through hierarchical learning, and then the invariance features of the target in the frequency domain are obtained by using the Fourier transform in polar coordinates. Finally, the features in the two-dimensional domains are compactically fused to obtain the multi-dimensional representation of the target features. In order to better adapt to the differences brought by SAR images collected by different sensors, Zhao et al. (2022) proposed an adaptive learning strategy based on the adversarial domain. Considering the different polarization modes and scattering intensity of SAR images, in order to realize the alignment of instance-level objects and pixel-level features between different domains (different sensor images), the concept of entropy is introduced as a feature weight coefficient to distinguish regions with different entropy. Since the entropy of the uniform region in SAR images is lower than that of the non-uniform region, adding entropy-based adversarial domain adaptive learning to different layers of the backbone network can effectively deal with the relationship between entropy and different receptive fields so that different domains can be aligned at the feature level as much as possible. At the same time, assigning different weights to regions with different entropy can help to distinguish the alignment results better. With the aim of distinguishing different instance-level target characteristics and make better alignment, the domain alignment compensation loss is constructed. In order to extract more precise feature information so that more uniquely representative example features can be accessed, the result of the highest score in the clustering is used to calculate the weight of the class. Zhou et al. (2023) added an edge semantic branch to solve problems such as confusion in edge detection caused by overlapping targets and used convolution of deeper and larger convolution kerns to expand the learning of context edge semantics and decouple the learned rich features, which is conducive to accurate localization of ship targets and prediction of detection frames. In addition, considering that the size of the receptive field extracted by CNN is limited, it is impossible to analyze the context from a global perspective. Therefore, a transformer framework is introduced to acquire global context features by using a multi-head attention mechanism, thus enhancing the remote analysis capability and achieving better detection and recognition effects for large-scale targets.

3 Context-Aware Network

In this section, we detail the overall architecture of the network and some other design-specific concepts and corresponding examples. The overall architecture of our approach is shown in Figure 2. Specifically, features are first extracted initially using CSPDarkNet53 as the backbone network. For the backbone network, our input goes through two convolutional layers to downsample the data to 1/4 of the input, where the activation function used in the convolutional layer is chosen to be the SiLU function. The SiLU function has a smoother curve as it approaches 0, controlling the output structure between 0 and 1 and achieving better results than ReLU in some applications. Then, the feature extraction method of YOLOv5 was adopted to obtain three effective feature layers with different resolutions and channel numbers through multiple C3 modules, and the three feature layers were input into the FPN network structure composed of LFRM and GCAM in series in parallel. The C3 module consists of three standard convolutional layers as well as multiple CSP Bottlenecks. The CSP Bottleneck mainly uses a residual structure, with one 1X1 convolution and one 3X3 convolution in the trunk, after which the residuals are left untouched and the inputs and outputs of the trunk are directly combined. The C3 module uses the CSPNet (Wang C. Y. et al., 2020) network structure, which still employs the residuals.

We capture multi-scale features through LFRM to better adapt to different scales of ship object information, thus obtaining a more representative feature map. Then, the long-range dependencies are captured by GCAM to enhance the feature representation of the target and suppress the interference of noise. The following subsections present detailed information.

3.1 LFRM

Since ship targets in SAR images in real applications may have different scales, some ships may be very large while others may be relatively small, making the detection process complicated. To address this problem, we designed the LFRM module as shown in Figure 3. The deep features $x = \{x_1 \ldots\ldots x_i\}$ obtained from the backbone network are computed in parallel by a 1×1 convolutional layer and three atrous convolutions with rates of 3, 6, and 12 to obtain convolutional features on multiple scales.

$$b_i = Atrous(x_i)$$
$$c_i = Conv_{1x1}(x_i)$$

After that, the feature maps b_i of each layer except the first one is sequentially fused with the feature maps b_{i-1} of the previous layer and activated by convolution to obtain new feature maps $\acute{b_i}$, which allows each layer to obtain a diversity of resceptive fields.

$$\acute{b_i} = Conv(b_i)$$

To better fuse the different scales of information, the four obtained feature maps are finally superimposed in the channel dimension using the Concat operation and then fed into the convolutional layer to obtain a new multi-scale feature map s_i.

$$s_i = Conv\left(Concat\left(\acute{b_i}, c_i\right)\right)$$

For the purpose of enhancing the generalization ability of the network, we improve ECA-Net (Wang Q. et al., 2020) by learning the correlation between channels and adaptively adjusting the weights of the channels to improve the performance of the network. As shown in the lower part of Figure 3, we first perform global maximum pooling and global average pooling operations on the feature map $x_i$to obtain two global feature descriptors, respectively, $m\in \mathbb{R}^{1\times1\times C}$, $a\in \mathbb{R}^{1\times1\times C}$, C indicates the number of channels.

$$m = GAP\left(x\right)$$
$$a = GMP(x)$$

The cross-channel information interaction is accomplished by two one-dimensional convolutions, respectively, and then the weight coefficients for each channel are calculated by SoftMax normalization. Where $w_i$is the result of channel interactions, w_i^j denotes the weights of the channel features, and y_i denotes the neighboring feature channels in a one-dimensional space. K is the

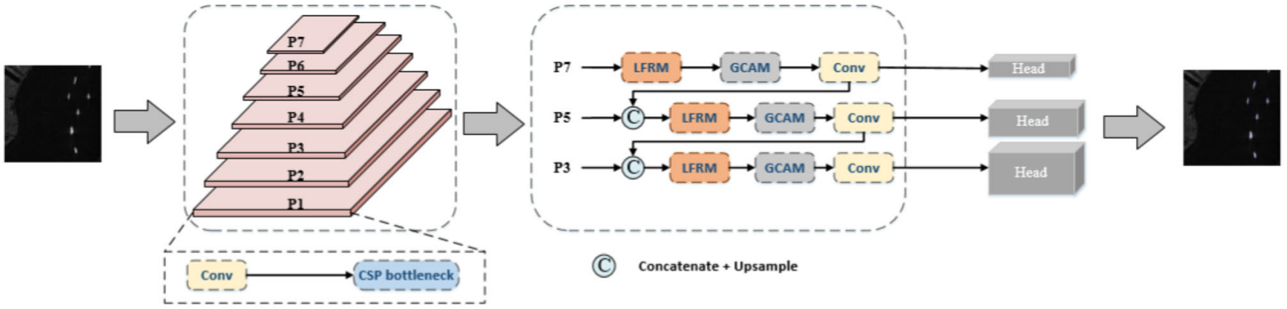


FIGURE 2
General framework of our method, where LFRM and GCAM are the proposed modules. The input image is first sent to the backbone to extract features, then passes through the FPN network structure consisting of LFRM and GCAM in series, and finally the detection results are output through the header. Where BCE loss is used for classification and objectivity and GIoU loss is used for regression.

result computed by the given formula, and i denotes the number of channels, $j \in \mathbb{R}^K$.

$$\omega_i = \left(\sum_{j=1}^{K} w_i^j y_i^j \right), y_i^j \in \Omega_i^K$$

Where the convolutional kernel size K is self-adapted by a function that allows layers with a larger number of channels to interact across channels more often. The adaptive convolutional kernel size is calculated as,

$$K = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{odd}$$

Which $\gamma = 2$, $b = 1$, $|t|_{odd}$ is the nearest odd number to t and C is the number of channel.

Finally, the results of the two different pooling branches are superimposed according to the channel dimension, and the weight coefficients for each channel are obtained using SoftMax normalization, andx_i is attentively weighted according to the channel dimension.

$$p = \sigma\left(Concat\left(\acute{m}, \acute{a}\right)\right) \cdot x$$

σ is SoftMax function, $\cdot$ is the element-wise product.

Finally, the multiscale feature s is overlaid with the feature map p after local cross-channel interaction to obtain the final LFRM output.

Since using only GAP to extract global features does not capture the detail information well, GM is added to enhance the grasp of details, and the two pooling branches complement each other to enhance the extraction of local semantic features.

3.2 GCAM

To obtain remote dependent features and thus global context information to enhance the ontological target characteristics and to remove the interference of complex background noise on the target, we design the GCAM module as shown in Figure 4, where we take the multi-scale information obtained from the LFRM module as an input to obtain the remote context information about the local features.

As shown in Figure 4, it given the output of the LFRM$P = \{P_1 \ldots\ldots P_i\}$ as input, $P_1 \in \mathbb{R}^{1\times C}$ is the feature vector at pixel i with C channel. The global context feature f_i is obtained by estimating the relationship between the current pixel and all pixels. After that, the weight coefficients are matrix multiplied with the local features to aggregate the contextual information (matrix multiplication is employed on the weight and local feature to aggregate contextual information).

$$f_i = \sum_{j=1}^{H\times W} \frac{e^{n(p_i)}}{\sum_{m=1}^{H\times W} e^{n(p_m)}} * p_j$$

Where $n(p_i) = W_k p_j$ and $n(p_m) = W_k p_m$ represent linear transform matrices, and W_k implements the 1 × 1 convolution.

With the aim of further extracting the channel dependencies while reducing the number of parameters and computational complexity, the acquisition of spatially distant effective features will be augmented by transformations, so we draw on the Non-local (Wang et al., 2018) method.

$$\bar{\acute{f}}_i = \theta * SiLU\left(\mathrm{LN}\left(\phi \cdot f_i\right)\right)$$

Both ϕ and θ are realized by a 1 × 1 convolution. And the normalization (LN) and SiLU activation layers are added after the first convolution to improve the generalization of the model. Finally, the transformed feature $\acute{f}_i$ is element-wise added to the multi-scale local features, yielding the GCAM output $\tilde{f}_i$ which aggregates global contextual features at each pixel.

$$\tilde{f}_i = \acute{f}_i + p_i$$

The GCAM module selectively acquires distant features for each pixel based on the correlation between spatially distant pixels, which enhances the modeling capability of feature representation and reduces background noise interference. Meanwhile, the module can be easily inserted into various network models to obtain global context information.

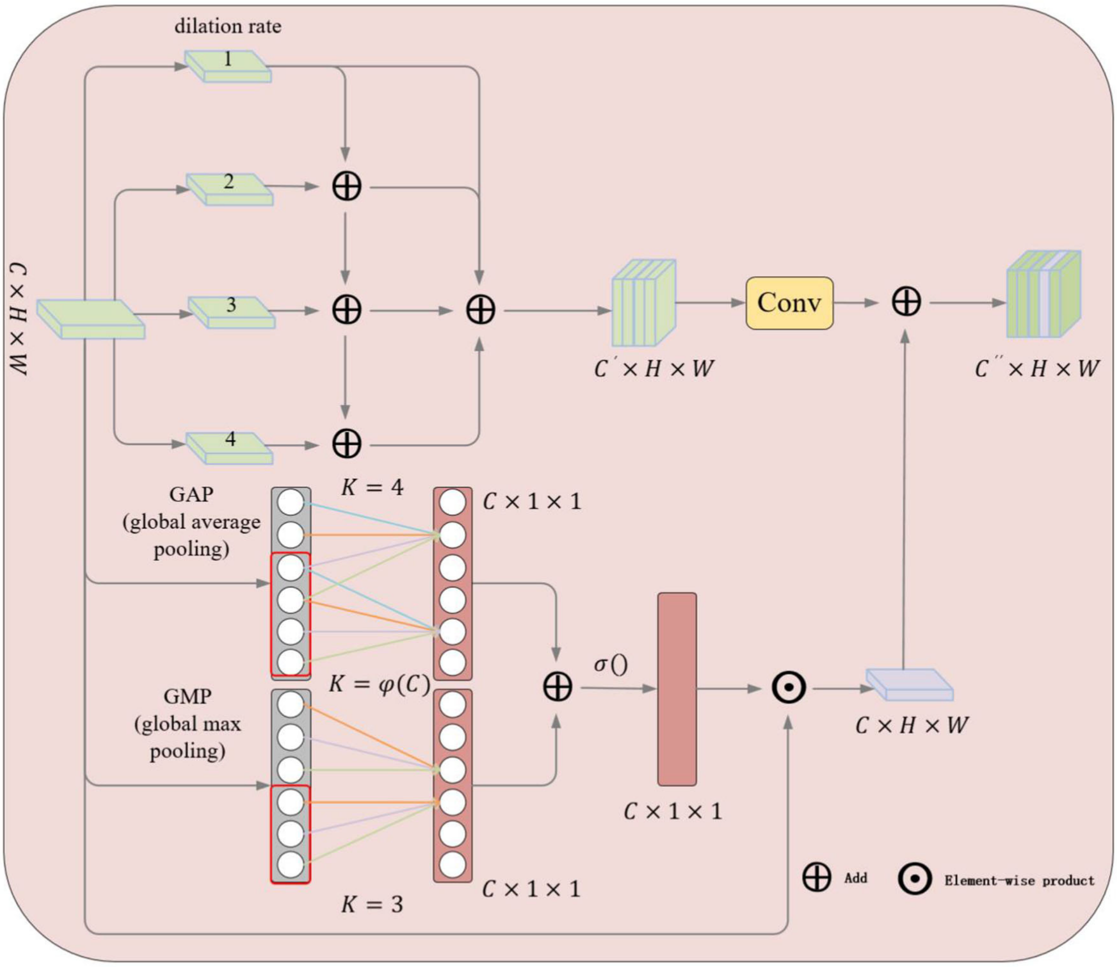


FIGURE 3
Illustration of the proposed LFRM. The upper half shows the extraction of multi-scale features using atrous convolution and the lower half shows the two-branch pooling channel attention mechanism.

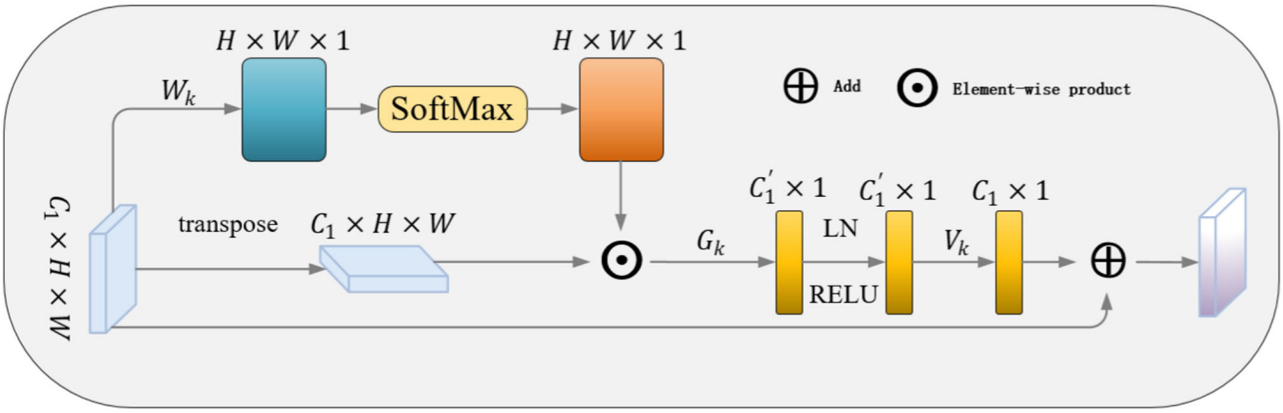


FIGURE 4
Illustration of the proposed GCAM.

4 Experimental results and analysis

In order to fully verify the validity of our proposed methods, we test them on three authoritative public data sets and compare them with several other advanced ones. In addition, to demonstrate the effectiveness of proposed LFRM and GCAM, we design ablation experiments to evaluate the validation. Finally, we provide a comprehensive analysis of the experimental results and time complexity.

4.1 Training configurations and datasets

All of our experiments are conducted on a GPU workstation equipped with NVIDIA RTX 3090 with 24 GB of video memory, and the operating systems are ubuntu21.0, CUDA (10.0) and cuDNN7.0. The language and framework used to build the model are python3.7 and pytorch1.1.0, respectively. For achieving fast convergence during training, with AdamW optimizer, we set the initial learning rate to 1e-3 and employ a cosine annealing strategy to adjust. Also, to ensure experimental fairness and consistency, all the methods involved in the experiments are trained and validated under the same data benchmark. The batch setting is 16 and the maximum number of iterations is 300 to find the best model parameters.

The loss function, which used for model training, consists of classification loss, confidence loss and regression localization loss. The former two chose the classical Cross Entropy (CE), while the latter adapts Complete-IoU (CIoU) Loss.

The Cross-Entropy Loss L_{CE} function expression is shown below, where $p\left(x_i\right)$ is the probability distribution of the true value, $q\left(x_i\right)$is the probability distribution of the predicted value, and **C** denotes the total number of categories.

$$L_{CE} = -\sum_{i=1}^{C} p\left(x_i\right) \ln\left(q\left(x_i\right)\right)$$

The CIOU loss L_{CIOU} function expression is shown below, where $\rho^2(b, b^{gt})$ represents the square of the distance between the center point of the prediction box and the center point of the real box. c represents the diagonal length of the smallest outer rectangle of the two rectangular boxes. α is the parameter used to do trade-offs, and v is the parameter used to measure aspect ratio consistency.

$$CIoU = IoU - (\frac{\rho^2(b, b^{gt})}{c^2} + \alpha\upsilon)$$
$$\upsilon = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2$$
$$\alpha = \frac{v}{(1 - IoU) + v}$$
$$L_{CIoU} = 1 - CIoU$$

The CIOU loss was chosen to normalize the coordinate scales to take advantage of the IOU and initially address the case where the IOU is zero.

To more fully evaluate the superiority of our methods, AP50 is used as the main evaluation metric, compared with currently popular methods. Specifically, PR curve is a curve drawn with precision P as the vertical coordinate and recall rate R as the horizontal coordinate. The higher the accuracy of the model, the higher the recall rate, the better the model performance, and the larger the area under the PR curve. AP50 Indicates the AP value when the IoU confidence score is 0.5. In addition, we use accuracy, recall, and F1 scores for a confidence threshold of 0.4. We also use FLOPs as an auxiliary evaluation metrics to test the efficiency of the model. The formula for calculating indicators is as follows:

$$Precision = \frac{TP}{TP + FP}$$
$$Recall = \frac{TP}{TP + FN}$$
$$Fl = \frac{2 \times Precision \times Recall}{Precision + Recall}$$
$$AP = \int_0^1 P(R)dR$$

4.2 Datasets

We evaluate our proposed methods on several public SAR ship datasets, including the SAR-Ship-Dataset (Wang et al., 2019), HRSID (Wei et al., 2020), and SSDD (Li et al., 2017) datasets. All of these datasets contain real scene images of various complex scenes ship targets of different sizes and dimensions. The SAR-Ship-Dataset (Wang et al., 2019) annotated by SAR experts, which uses 102 SAR images taken by the Gaofen-3 satellite and 108 SAR images taken by the Sentinel-1 satellite, containing 43,819 slices and 50,885 ship targets. The pixels in distance and orientation are 256. Finally, the data set is randomly divided into training set, verification set, and test set, with an image ratio of 7:2:1. HRSID (Wei et al., 2020) is a public data set used for the ship detection, semantic segmentation, and instance segmentation in high-resolution SAR images. It contains 5,604 high-resolution SAR ship images and 16,951 ship instances. The construction process draws on the COCO dataset and includes SAR images of different resolutions, polarization modes, sea states, sea areas, and ports. Its spatial resolution is 0.5–3 m. We follow the original dataset paper's delineation method. For the SSDD (Li et al., 2017) dataset is obtained by downloading publicly available SAR images from the Internet and cropping the target area into 1,160 pixels of size around 500 × 500 and manually labeling the ship target positions. We select images with image index suffixes 1 and 9 as the test set.

4.3 Results and analysis

4.3.1 SAR-ship-dataset

As shown in Table 1, our algorithms are experimentally compared with general-purpose object detection methods including Faster R-CNN (Ren et al., 2015), RetinaNet (Lin et al., 2017a), CenterNet (Zhou et al., 2019), YOLOv4 (Bochkovskiy et al., 2020), and YOLOv5, as well as SAR-specific ship detectors DAPN (Cui et al., 2019), CoAM+RFIM (Yang et al., 2021), and PPA-Net (Tang et al., 2023) on the SAR ship dataset (Wang et al., 2019). From the Table 1, it can be observed that our method exhibits

TABLE 1 Comparison of evaluation metrics of different methods on the SAR-SHIP dataset.

Method	Precision (%)	Recall (%)	F1 (%)	AP_{50} (%)
Faster R-CNN (Ren et al., 2015)	90.3	91.4	90.8	91.0
RetinaNet (Lin et al., 2017a)	84.5	93.3	88.7	93.8
CenterNet (Zhou et al., 2019)	84.6	93.5	88.8	95.0
DAPN (Cui et al., 2019)	89.9	90.7	90.3	90.6
YOLOv4 (Bochkovskiy et al., 2020)	85.7	92.7	89.1	93.2
YOLOv5	93.5	95.0	**94.9**	95.8
CoAM+RFIM (Yang et al., 2021)	93.7	95.3	94.5	96.0
PPA-Net (Tang et al., 2023)	93.5	95.5	94.7	96.1
Our	**93.8**	**96.1**	94.4	**96.3**

The best results are highlighted in bold.

TABLE 2 Comparison of evaluation metrics of different methods on the HRSID dataset.

Method	Precision (%)	Recall (%)	F1 (%)	AP_{50} (%)
Faster R-CNN (Ren et al., 2015)	88.8	77.5	82.8	78.2
RetinaNet (Lin et al., 2017a)	69.8	83.8	76.2	82.5
CenterNet (Zhou et al., 2019)	81.8	87.4	84.5	86.3
DAPN (Cui et al., 2019)	88.9	77.6	82.9	79.8
YOLOv4 (Bochkovskiy et al., 2020)	90.6	84.0	87.2	90.1
YOLOv5	92.4	89.3	91.2	92.9
CoAM+RFIM (Yang et al., 2021)	92.7	88.1	90.3	92.7
PPA-Net (Tang et al., 2023)	93.4	89.8	92.1	92.9
Our	**93.6**	**90.4**	**92.4**	**93.3**

The best results are highlighted in bold.

TABLE 3 Comparison of evaluation metrics of different methods on SSDD dataset.

Method	Precision (%)	Recall (%)	F1 (%)	AP_{50} (%)
Faster R-CNN (Ren et al., 2015)	90.9	87.6	89.2	88.3
RetinaNet (Lin et al., 2017a)	81.6	92.3	86.6	89.6
CenterNet (Zhou et al., 2019)	93.3	**94.5**	93.9	93.5
DAPN (Cui et al., 2019)	87.6	91.4	89.4	90.1
YOLOv4 (Bochkovskiy et al., 2020)	93.6	94.0	93.8	96.1
YOLOv5	94.0	92.4	92.7	95.3
CoAM+RFIM (Yang et al., 2021)	94.4	92.1	93.2	95.6
PPA-Net (Tang et al., 2023)	94.8	**94.5**	93.3	96.0
Our	94.2	93.9	**94.5**	**96.2**

The best results are highlighted in bold.

strong competitiveness. Our approaches achieve precision, recall, F1, and AP50 accuracy of 93.8, 96.1, 94.4, and 96.3%, respectively. Regarding AP50 accuracy, it outperforms the two-stage detector Faster R-CNN (Ren et al., 2015) in general object detection by 5.3%, and exceeds YOLOv4 (Bochkovskiy et al., 2020) and YOLOv5 (both are single-stage detectors) by 3.1 and 0.5%, respectively.

In addition, in comparison with SAR ship detection method DAPN (Cui et al., 2019), which primarily focuses on the scale issue of ship targets but neglects the interference and impact of noise in small targets within complex backgrounds, resulting in an AP50 accuracy of 90.6%, significantly lower than ours and other advanced SAR ship detection methods. Our approach also outperforms another anchor-free popular algorithm, CoAM+RFIM (Yang et al., 2021), by 0.3% in the AP50 metric. Despite the consideration of noise impact and the use of attention mechanisms to reduce noise effects, the latest SAR ship detection method PPA-Net (Tang et al., 2023) falls short due to relying solely on pooling operations to address multi-scale information, leading to significant information loss.

4.3.2 HRSID

The HRSID dataset exhibits a more complex image background and includes a greater number of densely packed small ship targets, posing higher challenges for algorithms and allowing for a better validation of our method's effectiveness in complex background and small target detection. As shown in Table 2, our method shows an improvement of ∼0.4–15.1% compared to state-of-the-art methods, benefiting from the proposed LFRM and GCAM. LFRM first extracts local multiscale information using multiple differently-sized receptive fields and then employs a dual-branch channel attention mechanism to facilitate local cross-channel information interaction between different scale features, alleviating the detection impact of scale variations.

Furthermore, GCAM, by capturing long-range dependencies, enhances target feature representation and suppresses noise interference, enabling effective target detection in SAR ship images with different complex backgrounds. Even when compared to the latest SAR ship detection algorithms CoAM+RFIM (Yang et al., 2021) and PPA-Net (Tang et al., 2023), our method outperforms them by 0.9, 2.3, 2.1, and 0.2% for Precision (%), Recall (%), F1 (%), and AP50 (%), respectively. Similarly, across all four detection accuracy metrics, our method surpasses other general object detection methods and achieves optimal results. In terms of AP50 (%), it outperforms Faster R-CNN (Ren et al., 2015), RetinaNet (Lin et al., 2017a), CenterNet (Zhou et al., 2019), YOLOv4 (Bochkovskiy et al., 2020), and YOLOv5 by 15.1, 10.8, 13.5, 3.2, and 0.4%, respectively.

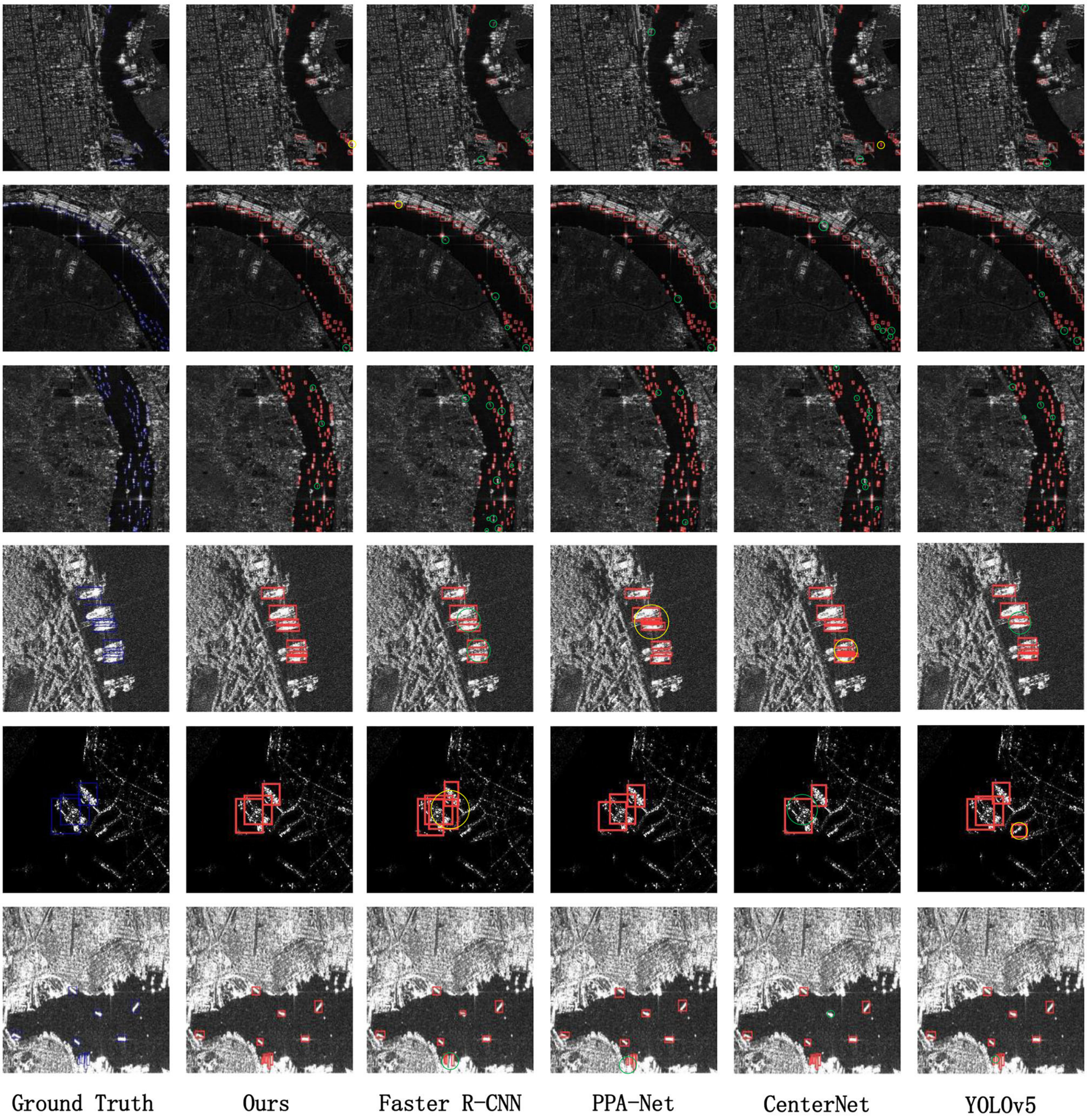


FIGURE 5
We have chosen to compare the detection results of different methods for complex backgrounds and multi-scale targets (especially small targets). The red box indicates the ground truth, and false alarms and missed detections are circled using yellow and green circles, respectively.

4.3.3 SSDD

As shown in Table 3, the experimental results on this dataset indicate that our method is competitive, although the Precision and Recall accuracies are slightly lower than YOLOv4 (Bochkovskiy et al., 2020), CoAM+RFIM (Yang et al., 2021), and PPA-Net (Tang et al., 2023). Furthermore, our algorithm outperforms other classical methods, including Faster R-CNN (Ren et al., 2015), RetinaNet (Lin et al., 2017a), CenterNet (Zhou et al., 2019), DAPN (Cui et al., 2019), and YOLOv5. In summary, our method achieves significant detection accuracy. Additionally, the detection results on multiple datasets validate the fine generalization capability of this method.

4.3.4 Visual results

To directly showcase the advanced detection results of our method, we visualize the detection outcomes on three different datasets. As illustrated in Figures 5–7, it is evident that our method performs exceptionally well in both complex background and various-sized ship targets, surpassing other approaches.

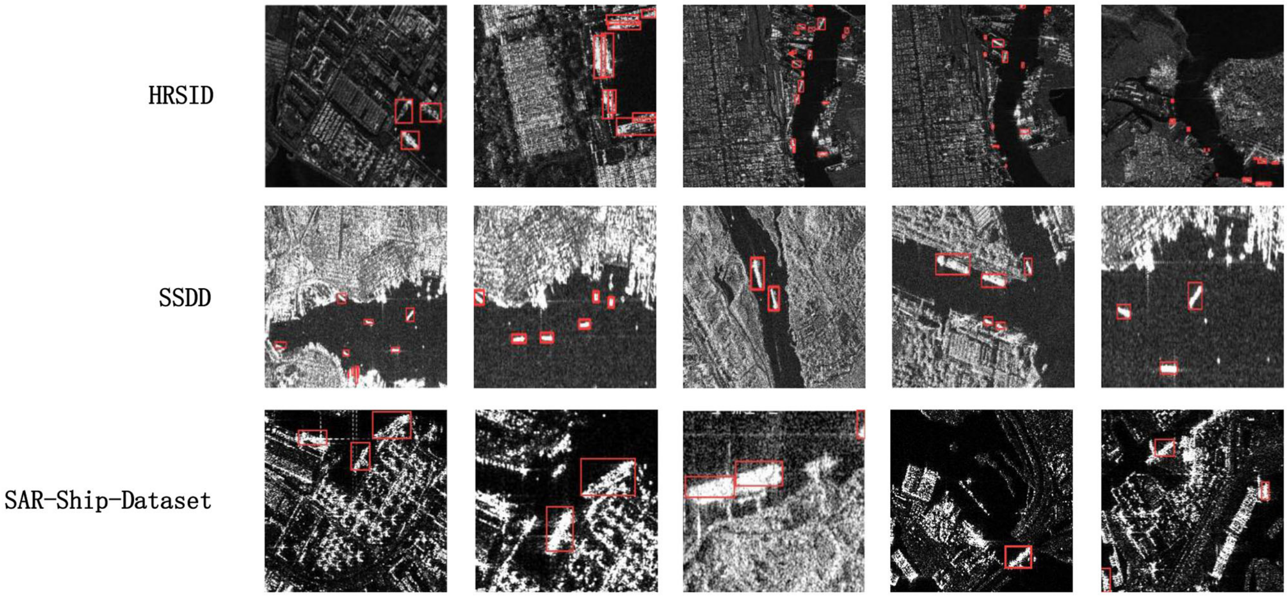


FIGURE 6
Plot of detection results for selected ships with complex backgrounds from HRSID, SSDD, and SAR-Ship-Dataset datasets for our method.

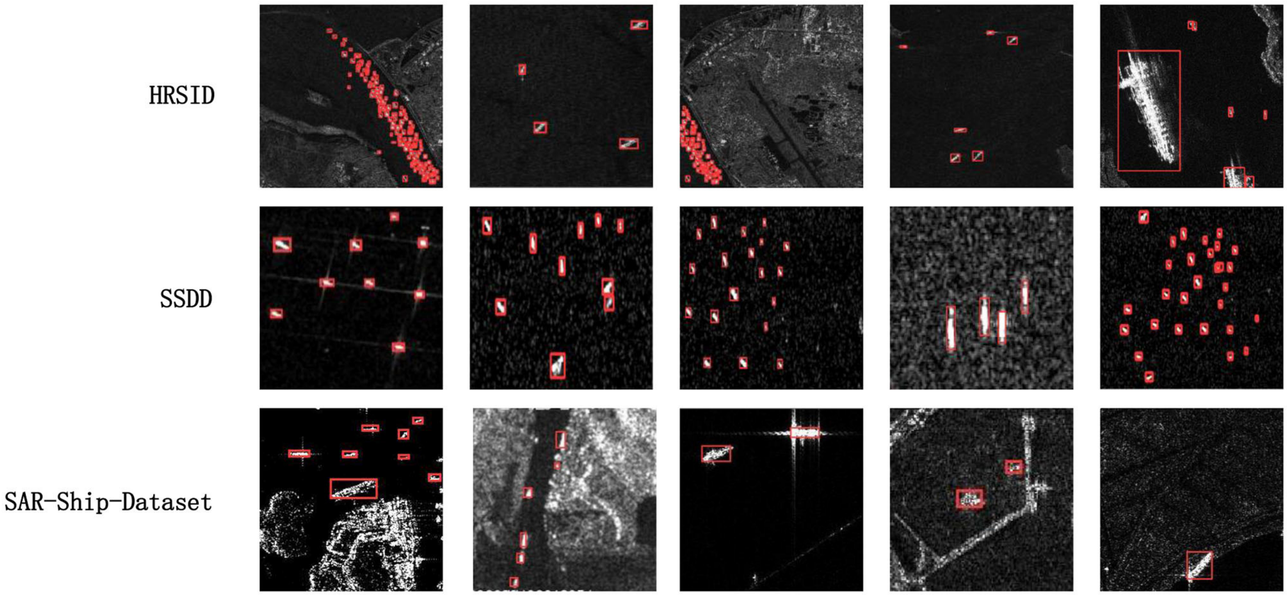


FIGURE 7
Our approach plots a selection of detection results with small targets and densely arranged ships in the HRSID, SSDD, and SAR-Ship-Dataset datasets.

Specifically, Figure 5 displays the detection results of our method and other approaches in SAR images with complex backgrounds and multiple-scale targets. It is noticeable that other methods exhibit instances of missed detections or false positives, while ours demonstrates good detection accuracy in both scenarios. Figure 6 presents the detection results of our method for ships with complex backgrounds. Figure 7 illustrates the results of detecting small target ships, consistent with our expectations that the LFRM module can effectively utilize multiple receptive fields of different sizes to extract local multiscale information, making the network more sensitive to small targets.

In summary, the visualization results intuitively reflect that our proposed method can accurately detect and identify ship targets in SAR images with complex backgrounds and various target sizes. Moreover, it demonstrates effective target detection across different datasets and diverse scenarios, offering better practical utility. However, our method exhibits some instances of missed detections and false positives in dense target detection, as shown in Figure 5, where our method displays a few missed detections in SAR images with densely packed ships, marked with green circles. This is attributed to our method solely considering the influences of multiscale targets and backgrounds, without accounting for potential feature overlap and misalignment that may arise when targets are densely arranged. Our current approach does not perform feature subdivision for overlapping targets, and we plan to address this in future work.

TABLE 4 Ablation experiments on the HRSID dataset.

LFRM	GCAM	AP_{50} (%)	Runtime (ms)
		91.1	9.1
✓		92.3	24.3
	✓	93.0	26.9
✓	✓	**93.3**	**28.1**

We validate the effectiveness of each component step by step. It displays the AP50 (%) and the Runtime (ms). The optimal metrics have been bolded. All scores are expressed in percentage (%).

TABLE 5 Ablation experiments on the HRSID dataset for the size selection of the convolutional region K in two-branch channel attention.

The coverage of K	AP_{50} (%)	Runtime (ms)
3	93.0	**19.7**
4	**93.3**	20.1
5	93.1	20.4
6	92.8	20.9

The bold values indicate the best results.

4.3.5 Ablation study

To evaluate the effectiveness of the components in our proposed Context-Aware Network, we conduct extensive ablation experiments on the HRSID (Wei et al., 2020) dataset. For LFRM, the results are shown in Table 4, where our proposed LFRM module improves the accuracy of AP50 from 91.1 to 92.3% compared to the benchmark level. As shown in Table 5, consistent with what we envisioned, LFRM uses multi-level atrous convolution to extract feature information at different scales hierarchically, and adopts residual linking to diversify the feature receptive field at each layer, better fusing the scale features. Combined with the dual-branch channel attention mechanism to realize local cross-channel interaction, it can enhance the ability to characterize the target and efficiently filter complex semantic information. The ablation experiments also demonstrate that LFRM is not only sensitive to scale information but also can mitigate complex background noise.

For GCAM, our proposed GCAM module improves the accuracy of AP50 from 91.1 to 93.0% compared to the benchmark level. Essentially, GCAM expands the sensory domain of the network by adaptively weighting features in different spaces and suppresses background noise interference by obtaining global contextual information based on the estimated long-range dependency. As shown in Figure 8, to show the effectiveness of our proposed module more directly, we visualize it by outputting a visual graph of the intermediate results. Finally, by combining our two modules in series, their AP50 accuracy can reach 93.3%, which shows that the LFRM and GCAM can effectively improve the SAR ship detection performance, and the interaction can further improve our network performance.

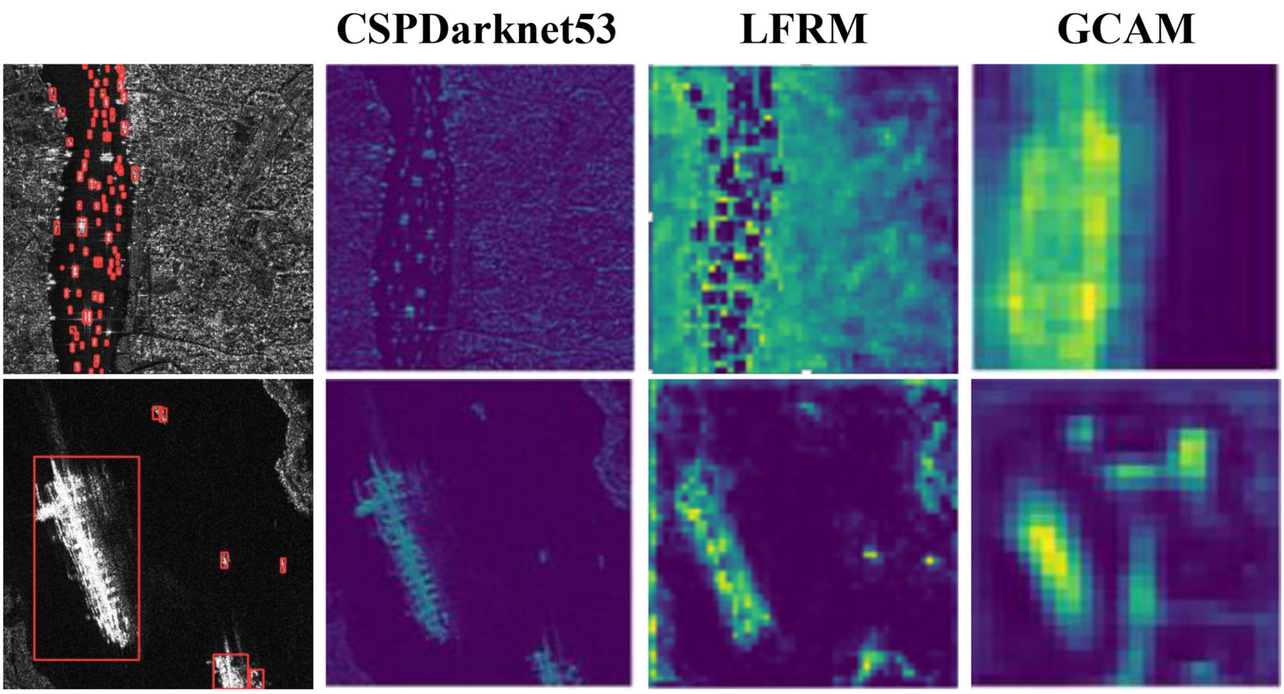


FIGURE 8
Visualization of the outputs of the different modules of the intermediate process tested by our method on the HRSID dataset.

To mitigate the impact of Batch Size on experimental results and determine the optimal Batch Size for training, we conduct ablation experiments with different Batch Size values. The experimental results are presented in Table 6. Notably, when the Batch Size reaches 16 and 32, the detection accuracy (AP50) both achieve the highest value of 93.3%. However, with a Batch Size of 8, the larger randomness introduced by the smaller Batch Size makes it challenging to converge, resulting in a lower classification accuracy of only 92.8%. When the Batch Size exceeds 32, there is a possibility of encountering local optima, leading to a decrease in accuracy to 92.9%. We exhaustively explored a range of Batch Size values in the ablation experiments to identify the most optimal Batch Size.

4.3.6 The complexity and speed of the network

We conduct a complexity analysis of the model, and the results are presented in Table 7. Ours has metrics of 28.1, 60.4, and 126.9 for Runtime, Params, and FLOPs, and although it is more complex to model with some other state-of-the-art methods such as YOLOv5, CoAM+RFIM (Yang et al., 2021) and PAA-Net (Tang et al., 2023), our method exhibits outstanding performance on the SAR-Ship-Dataset (Wang et al., 2019), HRSID (Wei et al., 2020), and SSDD (Li et al., 2017) datasets, delivering exceptional results while maintaining acceptable model sizes. The reason for the more complex model is that we use a more complex backbone network and GCAM in by calculating the correlation between each pixel and the other pixels, which imposes some network burden, but our method achieves a good balance for accuracy and speed.

TABLE 6 Ablation experiments were performed on HRSID data sets with different batch sizes.

Batch size	AP_{50} (%)
8	92.8
16	**93.3**
24	93.0
32	**93.3**
36	92.9

The bold values indicate the best results.

TABLE 7 Comparison of Runtime, Params size, and FLOPs for different models.

Method	Runtime (ms)	Params (M)	FLOPs (G)
Faster R-CNN (Ren et al., 2015)	56.1	60.1	181.9
RetinaNet (Lin et al., 2017a)	55.0	55.1	175.4
CenterNet (Zhou et al., 2019)	55.0	**20.2**	63.3
DAPN (Cui et al., 2019)	74.9	63.8	266.1
YOLOv4 (Bochkovskiy et al., 2020)	22.4	64.3	110.5
YOLOv5	**19.7**	27.6	**60.3**
CoAM+RFIM (Yang et al., 2021)	37.3	65.8	123.5
PPA-Net (Tang et al., 2023)	40.2	73,9	144.5
Our	28.1	70.4	126.9

The bold values indicate the best results.

5 Conclusion

To address the two challenges of various complex background interferences and multi-scale ship targets in SAR image ship detection tasks, we propose a context-aware one-stage SAR ship detection algorithm. To solve the problem of multi-scale ship target detection, we propose the LFRM module, which uses dilated convolutions with different ratios to obtain multi-scale features, and then uses average and maximum global pooling to interact the extracted information of different scales, enhancing its representation ability and sensitivity to scale, and achieving multi-scale ship detection. Furthermore, we also design the GCAM module to enhance the analysis of global context information and further suppress the interference of noise from complex backgrounds on targets. Extensive experiments have demonstrated that our method outperforms the latest methods in comprehensive performance. The method proposed in this paper can effectively cope with the interference of complex background noise and detect ship targets of different scales. However, there are still some missed detection issues for densely arranged targets. In future work, we will pay more attention to the detection of densely arranged small targets.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

CL: Conceptualization, Writing – review & editing. CY: Writing – original draft. HL: Data curation, Writing – review & editing. ZW: Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Fundamental Research Funds for the Central Universities (Grant Nos. NJ2022025, NP2022450, and NJ2023029) and the National Natural Science Foundation of China (Grant No. 61901504).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bochkovskiy, A., Wang, C. Y., and Liao, H. Y. M. (2020). Yolov4: optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934.* doi: 10.48550/arXiv.2004.10934

Chaudhary, Y., Mehta, M., Goel, N., Bhardwaj, P., Gupta, D., and Khanna, A. (2021). "YOLOv3 remote sensing SAR ship image detection," in *Data Analytics and Management: Proceedings of ICDAM* (Singapore: Springer), 519–531.

Chen, L. C., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587.* doi: 10.48550/arXiv.1706.05587

Cui, Z., Li, Q., Cao, Z., and Liu, N. (2019). Dense attention pyramid networks for multi-scale ship detection in SAR images. *IEEE Trans. Geosci. Remote Sens.* 57, 8983–8997. doi: 10.1109/TGRS.2019.2923988

Eldhuset, K. (1996). An automatic ship and ship wake detection system for spaceborne SAR images in coastal regions. *IEEE Trans. Geosci. Remote Sens.* 34, 1010–1019. doi: 10.1109/36.508418

Ghosh, S., Sivasankar, T., and Anand, G. (2021). Performance evaluation of multi-parametric synthetic aperture radar data for geological lineament extraction. *Int. J. Remote Sens.* 42, 2574–2593. doi: 10.1080/01431161.2020.1856963

Girshick, R. (2015). *Fast r-cnn[C]//Proceedings of the IEEE International Conference on Computer Vision* (Santiago, CA), 1440–1448.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Columbus, OH), 580–587.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pat. Anal. Machine Intell.* 37, 1904–1916. doi: 10.1109/TPAMI.2015.2389824

Iervolino, P., and Guida, R. (2017). A novel ship detector based on the generalized-likelihood ratio test for SAR imagery. *IEEE J. Select. Top. Appl. Earth Observ. Remote Sens.* 10, 3616–3630. doi: 10.1109/JSTARS.2017.2692820

Li, D., Liang, Q., Liu, H., Liu, Q., Liu, H., and Liao, G. (2021). A novel multidimensional domain deep learning network for SAR ship detection. *IEEE Trans. Geosci. Remote Sens.* 60, 1–13. doi: 10.1109/TGRS.2021.3062038

Li, J., Qu, C., and Shao, J. (2017). "Ship detection in SAR images based on an improved faster R-CNN," in *2017 SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA)*, ed H. Guo (Beijing: IEEE), 1–6.

Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017b). "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Venice; Honolulu, HI), 2117–2125.

Lin, T. Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017a). "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice; Honolulu, HI), 2980–2988.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., et al. (2016). "SSD: single shot multibox detector," in *Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14 2016, Proceedings, Part I 14* (Cham: Springer International Publishing), 21–37.

Ma, X., Hou, S., Wang, Y., Wang, J., and Wang, H. (2022). Multiscale and dense ship detection in SAR images based on key-point estimation and attention mechanism. *IEEE Trans. Geosci. Remote Sens.* 60, 1–11. doi: 10.1109/TGRS.2022.3225438

Mateus, P., Nico, G., Tomé, R., Catalão, J., and Miranda, P. M. (2012). Experimental study on the atmospheric delay based on GPS, SAR interferometry, and numerical weather model data. *IEEE Trans. Geosci. Remote Sens.* 51, 6–11. doi: 10.1109/TGRS.2012.2200901

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). "You only look once: unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 779–788.

Redmon, J., and Farhadi, A. (2017). "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 7263–7271.

Redmon, J., and Farhadi, A. (2018). YOLOv3: an incremental improvement. *arXiv preprint arXiv:1804.02767.* doi: 10.48550/arXiv.1804.02767

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: towards real-time object detection with region proposal networks. *Adv. Neural Inform. Process. Syst.* 2015, 28. doi: 10.48550/arXiv.1506.01497

Renga, A., Graziano, M. D., and Moccia, A. (2018). Segmentation of marine SAR images by sublook analysis and application to sea traffic monitoring. *IEEE Trans. Geosci. Remote Sens.* 57, 1463–1477. doi: 10.1109/TGRS.2018.2866934

Rohling, H. (1983). Radar CFAR thresholding in clutter and multiple target situations. *IEEE Trans. Aerospace Electr. Syst.* 4, 608–621. doi: 10.1109/TAES.1983.309350

Schwegmann, C. P., Kleynhans, W., and Salmon, B. P. (2016). Synthetic aperture radar ship detection using Haar-like features. *IEEE Geosci. Remote Sens. Lett.* 14, 154–158. doi: 10.1109/LGRS.2016.2631638

Sun, Q., Liu, M., Chen, S., Lu, F., and Xing, M. (2023). Ship detection in SAR images based on multi-level superpixel segmentation and fuzzy fusion. *IEEE Trans. Geosci. Remote Sens.* 2023, 3266373. doi: 10.1109/TGRS.2023.3266373

Tang, G., Zhao, H., Claramunt, C., Zhu, W., Wang, S., Wang, Y., et al. (2023). PPA-Net: pyramid pooling attention network for multi-scale ship detection in SAR images. *Remote Sens.* 15, 2855. doi: 10.3390/rs15112855

Uijlings, J. R., Van De Sande, K. E., Gevers, T., and Smeulders, A. W. (2013). Selective search for object recognition. *Int. J. Comput. Vis.* 104, 154–171. doi: 10.1007/s11263-013-0620-5

Wang, C., Jiang, S., Zhang, H., Wu, F., and Zhang, B. (2013). Ship detection for high-resolution SAR images based on feature analysis. *IEEE Geosci. Remote Sens. Lett.* 11, 119–123. doi: 10.1109/LGRS.2013.2248118

Wang, C. Y., Liao, H. Y. M., Wu, Y. H., Chen, P. Y., Hsieh, J. W., and Yeh, I. H. (2020). "CSPNet: a new backbone that can enhance learning capability of CNN," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (Seattle, WA), 390–391.

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., and Hu, Q. (2020). "ECA-Net: efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA), 11534–11542.

Wang, X., Girshick, R., Gupta, A., and He, K. (2018). "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 7794–7803.

Wang, Y., Wang, C., Zhang, H., Dong, Y., and Wei, S. (2019). A SAR dataset of ship detection for deep learning under complex backgrounds. *Remote Sens.* 11:765. doi: 10.3390/rs11070765

Wei, S., Zeng, X., Qu, Q., Wang, M., Su, H., and Shi, J. (2020). HRSID: a high-resolution SAR images dataset for ship detection and instance segmentation. *IEEE Access* 8, 120234–120254. doi: 10.1109/ACCESS.2020.3005861

Xu, X., Zhang, X., Shao, Z., Shi, J., Wei, S., Zhang, T., et al. (2022b). A group-wise feature enhancement-and-fusion network with dual-polarization feature enrichment for SAR ship detection. *Remote Sens.* 14:5276. doi: 10.3390/rs14205276

Xu, X., Zhang, X., and Zhang, T. (2022a). Lite-yolov5: a lightweight deep learning detector for on-board ship detection in large-scene sentinel-1 sar images. *Remote Sens.* 14:1018. doi: 10.3390/rs14041018

Yang, X., Sun, H., Fu, K., Yang, J., Sun, X., Yan, M., et al. (2018). Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks. *Remote Sens.* 10:132. doi: 10.3390/rs10010132

Yang, X., Zhang, X., Wang, N., and Gao, X. (2021). A robust one-stage detector for multiscale ship detection with complex background in massive SAR images. *IEEE Trans. Geosci. Remote Sens.* 60, 1–12. doi: 10.1109/TGRS.2021.3128060

Zhang, C., Yang, C., Cheng, K., Guan, N., Dong, H., and Deng, B. (2022). MSIF: multisize inference fusion-based false alarm elimination for ship detection in large-scale SAR images. *IEEE Trans. Geosci. Remote Sens.* 60, 1–11. doi: 10.1109/TGRS.2022.3159035

Zhang, T., and Zhang, X. (2019). High-speed ship detection in SAR images based on a grid convolutional neural network. *Remote Sens.* 11:1206. doi: 10.3390/rs11101206

Zhang, T., and Zhang, X. (2020). ShipDeNet-20: an only 20 convolution layers and< 1-MB lightweight SAR ship detector. *IEEE Geosci. Remote Sens. Lett.* 18, 1234–1238. doi: 10.1109/LGRS.2020.2993899

Zhang, T., and Zhang, X. (2022a). A mask attention interaction and scale enhancement network for SAR ship instance segmentation. *IEEE Geosci. Remote Sens. Lett.* 19, 1–5. doi: 10.1109/TGRS.2022.3225226

Zhang, T., and Zhang, X. (2022b). A full-level context squeeze-and-excitation ROI extractor for SAR ship instance segmentation. *IEEE Geosci. Remote Sens. Lett.* 19, 1–5. doi: 10.1109/TGRS.2022.3226570

Zhang, T., Zhang, X., and Ke, X. (2021a). Quad-FPN: a novel quad feature pyramid network for SAR ship detection. *Remote Sens.* 13:2771. doi: 10.3390/rs13142771

Zhang, T., Zhang, X., Ke, X., Zhan, X., Shi, J., Wei, S., et al. (2020c). LS-SSDD-v1. 0: a deep learning dataset dedicated to small ship detection from large-scale Sentinel-1 SAR images. *Remote Sens.* 12:2997. doi: 10.3390/rs12182997

Zhang, T., Zhang, X., Liu, C., Shi, J., Wei, S., Ahmad, I., et al. (2021b). Balance learning for ship detection from synthetic aperture radar remote sensing imagery. *ISPRS J. Photogrammetry Remote Sens.* 182, 190–207. doi: 10.1016/j.isprsjprs.2021.10.010

Zhang, T., Zhang, X., Shi, J., and Wei, S. (2019). Depthwise separable convolution neural network for high-speed SAR ship detection. *Remote Sens.* 11:2483. doi: 10.3390/rs11212483

Zhang, T., Zhang, X., Shi, J., and Wei, S. (2020a). HyperLi-Net: a hyper-light deep learning network for high-accurate and high-speed ship detection from synthetic aperture radar imagery. *ISPRS J. Photogram. Remote Sens.* 167, 123–153. doi: 10.1016/j.isprsjprs.2020.05.016

Zhang, T., Zhang, X., Shi, J., Wei, S., Wang, J., Li, J., et al. (2020b). Balance scene learning mechanism for offshore and inshore ship detection in SAR images. *IEEE Geosci. Remote Sens. Lett.* 19, 1–5. doi: 10.1109/TGRS.2020.3038405

Zhao, S., Zhang, Z., Guo, W., and Luo, Y. (2022). An automatic ship detection method adapting to different satellites SAR images with feature alignment and compensation loss. *IEEE Trans. Geosci. Remote Sens.* 60, 1–17. doi: 10.1109/TGRS.2022.3223036

Zhao, Y., Zhao, L., Xiong, B., and Kuang, G. (2020). Attention receptive pyramid network for ship detection in SAR images. *IEEE J. Select. Top. Appl. Earth Observ. Remote Sens.* 13, 2738–2756. doi: 10.1109/JSTARS.2020.2997081

Zhou, X., Wang, D., and Krähenbühl, P. (2019). Objects as points. *arXiv preprint arXiv:1904.07850*. doi: 10.48550/arXiv.1904.07850

Zhou, Y., Zhang, F., Yin, Q., Ma, F., and Zhang, F. (2023). Inshore dense ship detection in SAR images based on edge semantic decoupling and transformer. *IEEE J. Select. Top. Appl. Earth Observ. Remote Sens.* 2023:3277013. doi: 10.1109/JSTARS.2023.3277013

Zhu, M., Hu, G., Zhou, H., and Wang, S. (2022). Multiscale ship detection method in SAR images based on information compensation and feature enhancement. *IEEE Trans. Geosci. Remote Sens.* 60, 1–13. doi: 10.1109/LGRS.2022.3227251

Zhu, X., Lyu, S., Wang, X., and Zhao, Q. (2021). "TPH-YOLOv5: improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Montreal), 2778–2788.

frontiers | Frontiers in Neurorobotics

TYPE Original Research
PUBLISHED 20 March 2024
DOI 10.3389/fnbot.2024.1380430

Check for updates

OPEN ACCESS

EDITED BY
Predrag S. Stanimirovic,
University of Niš, Serbia

REVIEWED BY
Liangming Chen,
Chinese Academy of Sciences (CAS), China
Yang Shi,
Yangzhou University, China

*CORRESPONDENCE
Zhengtai Xie
✉ xzt1xzt@vip.qq.com

RECEIVED 01 February 2024
ACCEPTED 22 February 2024
PUBLISHED 20 March 2024

CITATION
Wen L and Xie Z (2024) A data-driven acceleration-level scheme for image-based visual servoing of manipulators with unknown structure. *Front. Neurorobot.* 18:1380430. doi: 10.3389/fnbot.2024.1380430

A data-driven acceleration-level scheme for image-based visual servoing of manipulators with unknown structure

Liuyi Wen[1,2] and Zhengtai Xie[1,3]*

[1]The State Key Laboratory of Tibetan Intelligent Information Processing and Application, Qinghai Normal University, Xining, China, [2]School of Arts, Lanzhou University, Lanzhou, China, [3]School of Information Science and Engineering, Lanzhou University, Lanzhou, China

The research on acceleration-level visual servoing of manipulators is crucial yet insufficient, which restricts the potential application range of visual servoing. To address this issue, this paper proposes a quadratic programming-based acceleration-level image-based visual servoing (AIVS) scheme, which considers joint constraints. Besides, aiming to address the unknown problems in visual servoing systems, a data-driven learning algorithm is proposed to facilitate estimating structural information. Building upon this foundation, a data-driven acceleration-level image-based visual servoing (DAIVS) scheme is proposed, integrating learning and control capabilities. Subsequently, a recurrent neural network (RNN) is developed to tackle the DAIVS scheme, followed by theoretical analyses substantiating its stability. Afterwards, simulations and experiments on a Franka Emika Panda manipulator with eye-in-hand structure and comparisons among the existing methods are provided. The obtained results demonstrate the feasibility and practicality of the proposed schemes and highlight the superior learning and control ability of the proposed RNN. This method is particularly well-suited for visual servoing applications of manipulators with unknown structure.

KEYWORDS

recurrent neural network (RNN), image-based visual servoing (IBVS), data-driven technology, acceleration level, learning and control

1 Introduction

Robots can accurately perform complex tasks and have become a vital driving force in industrial production (Agarwal and Akella, 2024). Among industrial robots, redundant robots, equipped with multiple degrees of freedom (DOFs), have gained significant recognition and favor due to their exceptional flexibility and automation capabilities (Tang and Zhang, 2022; Zheng et al., 2024). Therefore, numerous control schemes are designed to extend the application range of redundant robots, such as medical services (Zeng et al., 2024) and visual navigation (Wang et al., 2023). Furthermore, in these application scenarios, information on the external environment and the robot's status is acquired from various sensors, especially for the image capture of visual information (Jin et al., 2023). Therefore, unknown situations inevitably exist caused by sensor limitations, environmental variability, and robot modification, which hinder the evolution of robot applications. To address this issue, intelligent algorithms based on data-driven technology are exploited to process the acquired information and convert it into knowledge to drive the regular operation of the robot system (Na et al., 2021; Xie et al., 2022). Yang et al. (2019) construct a robot learning system by improving the adaptive ability of a robot with the information interaction between the robot and environment, which enhances the safety and reliability of robot applications in reality Peng et al. (2023). Li et al. (2019) investigate a model-free control method to cope with the unknown Jacobian problems inside the robot system. On this basis, the dynamic estimation method of robot parameters is researched in the study by Xie and Jin (2023). However, the aforementioned methods primarily operate at the joint velocity level and cannot directly applicable to robots driven by joint acceleration.

As a crucial robot application, visual servoing simulates the bionic system of human eyes, which can obtain information about real objects through optical devices, thus dynamically responding to a visible object. The fundamental task of visual servoing is to impose the error between the corresponding image feature and the desired static reference to approach zero (Zhu et al., 2022). According to the spatial position or image characteristics of the robot, the visual servoing system can be categorized into two types: position-based visual servoing (PBVS) system (Park et al., 2012), which utilizes 3-D position and orientation information to adjust the robot's state, and image-based visual servoing (IBVS) system, which utilizes 2-D image information for guidance (Van et al., 2018). Recently, the research on visual servoing has achieved many unexpected results (Hashimoto et al., 1991; Malis et al., 2010; Zhang et al., 2017; Liang et al., 2018). For instance, visual servoing is applied to bioinspired soft robots in the underwater environment with an adaptive control method, which extends the scope of visual servoing (Xu et al., 2019). Based on the neural network method, a resolution scheme for IBVS is developed at the velocity level. This enables the manipulator to accurately track fixed desired pixels, resulting in fast convergence (Zhang and Li, 2018). However, the aforementioned methods are difficult to deal with the emergence of unknown conditions, such as focal length change, robot abrasion, or parameter variation. This is because these methods rely on accurate structural information of the robot vision system. To tackle this challenge, this study focuses on data-driven control of visual servoing for robots with an unknown Jacobian matrix.

Neural networks have gained significant recognition as powerful tools for solving challenging problems, such as automatic drive (Jin et al., 2024), mechanism control (Xu et al., 2023), and mathematical calculation (Zeng et al., 2003; Stanimirovic et al., 2015). In robot redundancy analysis, neural networks have shown superior performance. In recent decades, numerous control laws based on neural networks have been developed to harness the potential of redundant manipulators (Zhang and Li, 2023). One specific application of the neural network approach addresses the IBVS problem. In this context, the IBVS problem is formulated as a quadratic programming scheme and tackled using a recurrent neural network (RNN). The RNN drives the robot vision system's feature to rapidly converge toward the desired point (Zhang et al., 2017). Additionally, Li et al. (2020) investigate an inverse-free neural network technique to deal with the IBVS task, ensuring that the error approaches zero within a finite time while considering the manipulator's physical constraints.

Most control schemes accomplish the given task at the velocity level, especially for visual servoing applications (Hashimoto et al., 1991; Malis et al., 2010; Zhang et al., 2017; Liang et al., 2018; Van et al., 2018; Zhang and Li, 2018; Xu et al., 2019; Li et al., 2020). These velocity-level schemes control redundant robots via joint velocities. However, when confronted with acceleration or torque-driven robots, the velocity-level schemes exhibit limitations and cannot provide precise control. Furthermore, the velocity-level scheme may yield abrupt joint velocities that are impractical in real-world applications. Consequently, research on acceleration-level visual servoing for robot manipulators has become crucial (Keshmiri et al., 2014; Anwar et al., 2019). Motivated by the issues above, this study investigates the application of visual servoing in robots at the acceleration level. The technical route of this study is shown in Figure 1. As illustrated, the contributions of this study are shown as follows:

- An acceleration-level image-based visual servoing (AIVS) scheme is designed, taking into account multiple joint constraints.
- Considering potential unknown factors in the visual servoing system, a data-driven acceleration-level image-based visual servoing (DAIVS) scheme is developed, enabling simultaneous learning and control.
- RNNs are proposed to solve the AIVS scheme and DAIVS scheme, enabling visual servoing control of the manipulator. Theoretical analyses guarantee the stability of the RNNs.

In addition, the feasibility of the proposed schemes is demonstrated through simulative and experimental results conducted on a Franka Emika Panda manipulator with an eye-in-hand structure.

Before concluding this section, the remaining sections of the study are shown as follows. Section 2 presents the robot kinematics of visual servoing and introduces the data-driven learning algorithm, formulating the problem at the acceleration level. Section 3 constructs an AIVS scheme with the relevant RNN. Subsequently, considering the unknown factors, a DAIVS scheme and corresponding RNN are proposed, and theoretical analyses proved the learning and control ability of the RNN, as shown in Section 4. Section 5 provides abundant simulations and performance comparisons, embodying the proposed method's validity and superiority. Section 6 displays physical experiments on a real manipulator. Finally, Section 7 briefly concludes this study.

2 Preliminaries

In this section, the robot visual servoing kinematics and data-driven learning algorithm are introduced as the preliminaries. Note that this study specifically tackles the problem at the acceleration level.

2.1 Robot visual servoing kinematics

The forward kinematics, which contains the transformation between the joint angle $\phi(t) \in \mathbb{R}^m$ of a robot and the end-effector position and posture $\mathbf{s}(t) \in \mathbb{R}^6$, can be expressed as follows:

$$f(\phi(t)) = \mathbf{s}(t), \tag{1}$$

where $f(\cdot)$ is the non-linear mapping related to the structure of the robot. In view of strongly non-linear and redundant characteristics

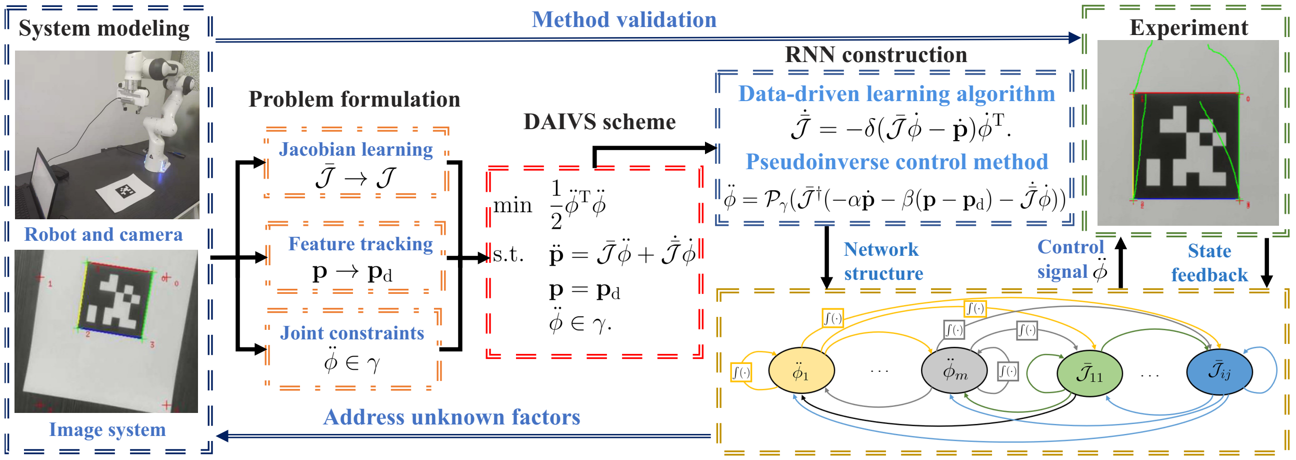

System modeling
Robot and camera
Image system
Method validation
Problem formulation
Jacobian learning
$\bar{\mathcal{J}} \to \mathcal{J}$
Feature tracking
$\mathbf{p} \to \mathbf{p}_{\rm d}$
Joint constraints
$\ddot{\phi} \in \gamma$
DAIVS scheme
$\min \ \frac{1}{2}\ddot{\phi}^{\rm T}\ddot{\phi}$
s.t. $\ddot{\mathbf{p}} = \bar{\mathcal{J}}\ddot{\phi} + \dot{\bar{\mathcal{J}}}\dot{\phi}$
$\mathbf{p} = \mathbf{p}_{\rm d}$
$\ddot{\phi} \in \gamma.$
RNN construction
Data-driven learning algorithm
$\dot{\bar{\mathcal{J}}} = -\delta(\bar{\mathcal{J}}\dot{\phi} - \dot{\mathbf{p}})\dot{\phi}^{\rm T}.$
Pseudoinverse control method
$\ddot{\phi} = \mathcal{P}_{\gamma}(\bar{\mathcal{J}}^{\dagger}(-\alpha\dot{\mathbf{p}} - \beta(\mathbf{p} - \mathbf{p}_{\rm d}) - \dot{\bar{\mathcal{J}}}\dot{\phi}))$
Experiment
Network structure
Control signal $\ddot{\phi}$
State feedback
$f(\cdot)$
$\ddot{\phi}_1$
$\ddot{\phi}_m$
$\bar{\mathcal{J}}_{11}$
$\bar{\mathcal{J}}_{ij}$
Address unknown factors

FIGURE 1
Technical route of this study.

of $f(\cdot)$, it is difficult to obtain the desired angle information directly from the desired end-effector information $\mathbf{s}_{\rm d}(t)$, i.e., $\mathbf{s}(t) = \mathbf{s}_{\rm d}(t)$. By taking the time derivative of both sides of Equation (1), one can deduce

$$J_{\rm ro}\dot{\phi}(t) = \dot{\mathbf{s}}(t), \tag{2}$$

where $\dot{\phi}(t)$ denotes the joint velocity; $\dot{\mathbf{s}}(t)$ covers the joint velocity and translational velocity of the end-effector; $J_{\rm ro} = \partial f(\phi(t))/\partial\phi(t) \in \mathbb{R}^{6\times m}$ stands for the robot Jacobian matrix. Owing to the physical properties of manipulators, output control signals based on design formulas and intelligent calculations may not be suitable for the normal operation of real robots. Therefore, to ensure the protection of the robot, it is crucial to take into account the following joint restrictions:

$$\begin{aligned} \phi^- &\le \phi \le \phi^+ \\ \dot{\phi}^- &\le \dot{\phi} \le \dot{\phi}^+ \\ \ddot{\phi}^- &\le \ddot{\phi} \le \ddot{\phi}^+, \end{aligned}$$

where ϕ^-, $\dot{\phi}^-$, and $\ddot{\phi}^-$ signify the lower bounds of joint angle, joint velocity, and joint acceleration, respectively; ϕ^+, $\dot{\phi}^+$ and $\ddot{\phi}^+$ denote the upper bounds of joint angle, joint velocity, and joint acceleration, respectively. Utilizing the special conversion techniques (Zhang and Zhang, 2012; Xie et al., 2022), the joint restrictions would be integrated into the acceleration level as $\ddot{\phi} \in \gamma$, where $\gamma = \{g \in \mathbb{R}^m, \gamma^- \le g \le \gamma^+\}$ is the safe range of joints with γ^- and γ^+ denoting the lower bound and upper bound of γ, respectively. In detail, the i-th elements of γ^- and γ^+ are designed as

$$\begin{aligned} \gamma_i^- &= \max\{\mu(\phi_i^- + \theta_i - \phi_i), \nu(\dot{\phi}_i^- - \dot{\phi}_i), \ddot{\phi}_i^-\} \\ \gamma_i^+ &= \min\{\mu(\phi_i^+ - \theta_i - \phi_i), \nu(\dot{\phi}_i^+ - \dot{\phi}_i), \ddot{\phi}_i^+\}, \end{aligned}$$

where $i = 1, 2, 3, \cdots, m$; $\mu > 0$ and $\nu > 0$ are designed to select the feasible region for different levels; θ_i is the margin to ensure that the acceleration has a sufficiently large feasible region (Xie et al., 2022). Then, a brief introduction to the visual servoing system is presented as follows. Regarding visual servoing tasks, the number of features determines the complexity of a visual servoing system. Simply considering a visual servoing system with one feature, a miniature camera is mounted on the end-effector of the manipulator and moves with the end-effector. Figure 2 illustrates the geometric transformation in different coordinate systems. Three-dimensional space with $O_{\rm ca}$ as the original point and $[X,\ Y,\ Z]$ as the coordinate axis is called the camera system with the internal coordinate point $\mathbf{q} = [x,\ y,\ z]^{\rm T}$. Relatively, with $O_{\rm im}$ as the center point, the image system is the two-dimensional space with the projection pixel point of $\mathbf{q}$ being $[p_x, p_y]^{\rm T}$ and the pixel coordinate being $\mathbf{p} = [p_u, p_v]^{\rm T} \in \mathbb{R}^2$. According to the similar triangle, it can be readily obtained in the study by Zhang et al. (2017) and Zhang and Li (2018):

$$\begin{bmatrix} p_x \\ p_y \end{bmatrix} = \frac{l}{z}\begin{bmatrix} x \\ y \end{bmatrix} \tag{3}$$

and

$$\begin{aligned} p_u &= u_0 + a_x p_x \\ p_v &= v_0 + a_y p_y, \end{aligned} \tag{4}$$

with l standing for the focal length of the camera; u_0 and v_0 denoting the pixel coordinate of principle point; and $[a_x,\ a_y]^{\rm T}$ standing for the conversion scale. Based on Equations (3, 4), the image Jacobian matrix $J_{\rm im}(\mathbf{p}, z) \in \mathbb{R}^{2\times 6}$ is defined using the following relationship (Liang et al., 2018):

$$J_{\rm im}(\mathbf{p}, z)\dot{\mathbf{s}} = \dot{\mathbf{p}}, \tag{5}$$

where $\dot{\mathbf{p}}$ stands for the movement velocity of the pixel coordinate and

$$J_{\rm im}(\mathbf{p}, z) = H\begin{bmatrix} -\frac{l}{z} & 0 & \frac{lp_x}{z} & \frac{p_x p_y}{l} & -\frac{p_x^2 + l^2}{l} & p_y \\ 0 & -\frac{l}{z} & \frac{p_y}{z} & -\frac{p_y^2 + l^2}{l} & -\frac{p_x p_y}{l} & -p_x \end{bmatrix},$$

with

$$p_x = \frac{p_u - u_0}{a_x},\quad p_y = \frac{p_v - v_0}{a_y},\quad H = \begin{bmatrix} a_x & 0 \\ 0 & a_y \end{bmatrix}.$$

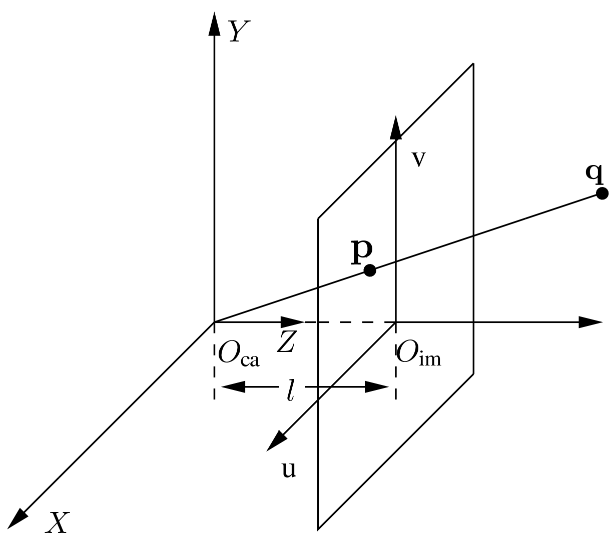


FIGURE 2
Geometric schematic of the camera system.

For the sake of convenience, Equations (2, 5) can be combined as follows:

$$\mathcal{J}\dot{\phi} = \dot{\mathbf{p}}, \tag{6}$$

with $\mathcal{J} = J_{\rm im}(\mathbf{p}, z)J_{\rm ro} \in \mathbb{R}^{2\times m}$ defined as the visual Jacobian matrix. The relationship between joint space and image space is established directly by Equation (6) at the velocity level. Taking the time derivatives of both sides of Equation (6) generates

$$\dot{\mathcal{J}}\dot{\phi} + \mathcal{J}\ddot{\phi} = \ddot{\mathbf{p}}, \tag{7}$$

where $\dot{\mathcal{J}}$ is the time derivative of $\mathcal{J}$; $\ddot{\phi}$ denotes the joint acceleration; and $\ddot{\mathbf{p}}$ stands for the movement acceleration of the pixel coordinate. When it comes to a complicated situation with more features, the above analyses still hold under the requirements of appropriate dimensions. It is worth noting that a single feature is analyzed as an example for simple illustration. When the number of features increases, the principle of coordinate transformation remains unchanged along with the increase in dimension.

2.2 Data-driven learning algorithm

However, unknown conditions may exist in the robot visual servoing system, such as focal length changes or robot modifications. In this regard, it could not control the robot accurately to execute the IBVS task based on $\dot{\mathcal{J}}$. Hence, motivated by this issue, a data-driven learning algorithm is designed as follows. To begin with, a virtual IBVS system is established, incorporating the virtual visual Jacobian matrix $\bar{\mathcal{J}} \in \mathbb{R}^{2\times m}$ and the following relationship:

$$\bar{\mathcal{J}}\dot{\phi} = \dot{\bar{\mathbf{p}}},$$

where $\dot{\bar{\mathbf{p}}} \in \mathbb{R}^2$ is the virtual pixel velocity determined by the virtual robot and $\dot{\phi}$ is the joint velocity measured in real time from the robot. Beyond dispute, the goal of the data-driven learning algorithm is to guarantee that $\dot{\bar{\mathbf{p}}}$ can rapidly converge to the real pixel velocity $\dot{\mathbf{p}}$. Thereout, an error function is devised as $\ell = ||\dot{\bar{\mathbf{p}}} - \dot{\mathbf{p}}||_2^2/2$, where $||\cdot||_2$ is the Euclidean norm of a vector. On the basis of the gradient descent method (Stanimirovic et al., 2015) to minimize the error function along the negative gradient direction, one can get

$$\dot{\bar{\mathcal{J}}} = -\delta\frac{\partial \ell}{\partial \bar{J}} = -\delta(\bar{\mathcal{J}}\dot{\phi} - \dot{\mathbf{p}})\dot{\phi}^{\rm T}, \tag{8}$$

where $\dot{\bar{\mathcal{J}}}$ is the time derivation of $\bar{\mathcal{J}}$; $\delta > 0$ denotes the coefficient that controls the convergence rate. Hereinafter, $\dot{\bar{\mathcal{J}}}$ and $\bar{\mathcal{J}}$ are used to replace the calibrated parameter $\dot{\mathcal{J}}$ and $\mathcal{J}$ to deal with the unknown situations. This method directly explores the relationship between joint space and image space without the utilization of $\dot{\mathcal{J}}$ and $\mathcal{J}$. It is worth highlighting that Equation (8) does not involve real structural information and estimates structural information from the joint velocity $\dot{\phi}$ and velocity of the pixel coordinate $\dot{\mathbf{p}}$ measured by sensors, which belongs to the core idea of the data-driven learning algorithm.

3 Acceleration-level IBVS solution

In this section, an AIVS scheme is proposed with joint constraints considered. Subsequently, we propose a corresponding RNN and provide theoretical analyses. Note that the presented method requires an accurate visual Jacobian matrix.

3.1 AIVS scheme

It is worth pointing out that there are few acceleration-level robot control schemes for dealing with IBVS problems. None of the existing acceleration-level solutions take joint constraints into account (Keshmiri et al., 2014; Anwar et al., 2019). In this regard, considering joint constraints, acceleration control, and visual servoing kinematics, the AIVS scheme is constructed as a quadratic programming problem, taking the following form:

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\ddot{\phi}^{\rm T}\ddot{\phi} && \text{(9a)}\\ &\text{subject to} && \ddot{\mathbf{p}} = \mathcal{J}\ddot{\phi} + \dot{\mathcal{J}}\dot{\phi} && \text{(9b)}\\ & && \mathbf{p} = \mathbf{p}_{\rm d} && \text{(9c)}\\ & && \ddot{\phi} \in \gamma, && \text{(9d)} \end{aligned}$$

where $\mathbf{p}_{\rm d}$ denotes the desired pixel coordinate. As a result, the goal of AIVS scheme (9) is to make the end-effector track the desired pixel point. In addition, according to robot Jacobian matrix $J_{\rm ro}$ and the image Jacobian matrix $J_{\rm im}$, the visual Jacobian matrix $\mathcal{J}$ and its time derivative $\dot{\mathcal{J}}$ are determined by the structure and parameters of the robot and the parameter settings inside the camera. Hence, if there are any changes in the internal parameters or structures, leading to an unknown state, the accuracy of $\mathcal{J}$ and $\dot{\mathcal{J}}$ may be compromised, potentially leading to a decline in performance. In contrast to velocity-level visual servoing schemes (Hashimoto et al., 1991; Malis et al., 2010; Zhang et al., 2017; Liang et al., 2018; Van et al., 2018; Zhang and Li, 2018; Xu et al., 2019; Li et al., 2020), the proposed AIVS scheme (9) offers two advantages. First, it utilizes joint acceleration as the control signal, resulting in continuous joint velocities. This helps mitigate the issues associated with excessive and discontinuous joint velocities. Second, AIVS scheme (9) takes into account the equality and inequality constraints at the acceleration level. This allows for a more comprehensive consideration of constraints, expanding the range of applications.

3.2 RNN solution and theoretical analysis

For the AIVS scheme (9), the pseudoinverse method is applied to generate the relevant RNN solution (Cigliano et al., 2015; Li et al., 2020). Primarily, as reported in the study by Zhang and Zhang (2012) and Xie et al. (2022), one can readily extend pixel coordinate error $\mathbf{p}-\mathbf{p}_{\mathrm{d}}$ into the acceleration level by neural dynamics method (Liufu et al., 2024) as

$$\ddot{\mathbf{p}}-\ddot{\mathbf{p}}_{\mathrm{d}}=-\alpha(\dot{\mathbf{p}}-\dot{\mathbf{p}}_{\mathrm{d}})-\beta(\mathbf{p}-\mathbf{p}_{\mathrm{d}}), \qquad (10)$$

where the design parameter $\alpha > 0$ and $\beta > 0$; $\dot{\mathbf{p}}_{\mathrm{d}}$ and $\ddot{\mathbf{p}}_{\mathrm{d}}$ are the desired velocity and the desired acceleration of the pixel coordinates, respectively. It is worth pointing out that the desired pixel coordinates $\mathbf{p}_{\mathrm{d}}$ is a constant, thus $\dot{\mathbf{p}}_{\mathrm{d}}=\ddot{\mathbf{p}}_{\mathrm{d}}=0$. As a result, Equation (10) can be rearranged as

$$\ddot{\mathbf{p}}=-\alpha\dot{\mathbf{p}}-\beta(\mathbf{p}-\mathbf{p}_{\mathrm{d}}). \qquad (11)$$

Substituting Equation (11) into Equation (9b), it could be obtained:

$$\dot{\mathcal{J}}\dot{\phi}+\mathcal{J}\ddot{\phi}=-\alpha\dot{\mathbf{p}}-\beta(\mathbf{p}-\mathbf{p}_{\mathrm{d}}).$$

In light of the pseudoinverse method, the joint acceleration can be minimized with the following formula:

$$\ddot{\phi}=\mathcal{J}^{\dagger}(-\alpha\dot{\mathbf{p}}-\beta(\mathbf{p}-\mathbf{p}_{\mathrm{d}})-\dot{\mathcal{J}}\dot{\phi}), \qquad (12)$$

where superscript † denotes the pseudoinverse operation of a matrix with $\mathcal{J}^{\dagger}=\mathcal{J}^{\mathrm{T}}(\mathcal{J}\mathcal{J}^{\mathrm{T}})^{-1}$. It is deserved to note that Equation (12) is employed in the study by Keshmiri et al. (2014) and Anwar et al. (2019) to generate the acceleration command for a manipulator. However, the research in the study by Keshmiri et al. (2014) and Anwar et al. (2019) does not consider joint constraints of the manipulator. To address this problem, the RNN corresponding to the AIVS scheme (9) is derived as

$$\ddot{\phi}=\mathcal{P}_{\gamma}(\mathcal{J}^{\dagger}(-\alpha\dot{\mathbf{p}}-\beta(\mathbf{p}-\mathbf{p}_{\mathrm{d}})-\dot{\mathcal{J}}\dot{\phi})), \qquad (13)$$

where projection function $\mathcal{P}_{\gamma}(c)=\mathrm{argmin}_{b\in\gamma}||b-c||_2$. Furthermore, theoretical analyses regarding the convergence of RNN (13) are presented as follows.

Theorem 1: The pixel error $\xi=\mathbf{p}-\mathbf{p}_{\mathrm{d}}$ driven by AIVS scheme (9) assisted with RNN (13) globally converges to a zero vector.

Proof: According to Equations (7, 13), one has

$$\ddot{\mathbf{p}}=\mathcal{J}\ddot{\phi}+\dot{\mathcal{J}}\dot{\phi}=\mathcal{J}\mathcal{P}_{\gamma}(\mathcal{J}^{\dagger}(-\alpha\dot{\mathbf{p}}-\beta(\mathbf{p}-\mathbf{p}_{\mathrm{d}})-\dot{\mathcal{J}}\dot{\phi}))+\dot{\mathcal{J}}\dot{\phi}.$$

Due to the fact that $\mathbf{p}_{\mathrm{d}}$ is a fixed feature, error function $\ddot{\xi}$ can be readily derived as

$$\ddot{\xi}=\mathcal{J}\mathcal{P}_{\gamma}(\mathcal{J}^{\dagger}(-\alpha\dot{\xi}-\beta\xi-\dot{\mathcal{J}}\dot{\phi}))+\dot{\mathcal{J}}\dot{\phi}.$$

By considering the projection function, a substitution matrix $\mathcal{H}$ is designed to replace $\mathcal{P}_{\gamma}(\cdot)$, leading to

$$\ddot{\xi}=\mathcal{J}\mathcal{H}\mathcal{J}^{\dagger}(-\alpha\dot{\xi}-\beta\xi-\dot{\mathcal{J}}\dot{\phi})+\dot{\mathcal{J}}\dot{\phi}, \qquad (14)$$

of which

$$\mathcal{H}=\begin{bmatrix} h_1 & 0 & \cdots & 0 \\ 0 & h_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h_m \end{bmatrix}\in\mathbb{R}^{m\times m}$$

and

$$h_i=\frac{\left(\mathcal{P}_{\gamma}(\mathcal{J}^{\dagger}(-\alpha\dot{\mathbf{p}}-\beta(\mathbf{p}-\mathbf{p}_{\mathrm{d}})-\dot{\mathcal{J}}\dot{\phi}))\right)_i}{\left(\mathcal{J}^{\dagger}(-\alpha\dot{\mathbf{p}}-\beta(\mathbf{p}-\mathbf{p}_{\mathrm{d}})-\dot{\mathcal{J}}\dot{\phi})\right)_i}\in(0,1].$$

By matrix decomposition, structural analyses of matrix $\mathcal{J}\mathcal{H}\mathcal{J}^{\dagger}=[a_{11},a_{12};a_{21},a_{22}\in\mathbb{R}^{2\times 2}$ are given as follows:

$$\mathcal{J}\mathcal{H}\mathcal{J}^{\dagger}=\mathcal{J}LL^{\mathrm{T}}\mathcal{J}^{\mathrm{T}}(\mathcal{J}\mathcal{J}^{\mathrm{T}})^{-1},$$

where $L=\sqrt{\mathcal{H}}$. In this regard, matrix $\mathcal{J}\mathcal{H}\mathcal{J}^{\dagger}$ can be viewed as the product of two positive definite matrices. It is evident that the eigenvalues of $\mathcal{J}\mathcal{H}\mathcal{J}^{\dagger}$ are greater than zero and $\det(\mathcal{J}\mathcal{H}\mathcal{J}^{\dagger})=\det(\mathcal{J}LL^{\mathrm{T}}\mathcal{J}^{\mathrm{T}})\det((\mathcal{J}\mathcal{J}^{\mathrm{T}})^{-1})>0$ with $\det(\cdot)$, denoting the determinant of a matrix. According to the properties of the diagonal elements of the matrix, it can be concluded that the diagonal elements of $\mathcal{J}\mathcal{H}\mathcal{J}^{\dagger}$ are greater than zero ($a_{11}>0$, $a_{22}>0$). Furthermore, Equation (12) can be rewritten as

$$\begin{bmatrix}\ddot{\xi}_1\\ \ddot{\xi}_2\end{bmatrix}=\begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix}\begin{bmatrix}-\alpha\dot{\xi}_1-\beta\delta\xi_1-(\dot{\mathcal{J}}\dot{\phi})_1\\ -\alpha\dot{\xi}_2-\beta\xi_2-(\dot{\mathcal{J}}\dot{\phi})_2\end{bmatrix}+\begin{bmatrix}(\dot{\mathcal{J}}\dot{\phi})_1\\ (\dot{\mathcal{J}}\dot{\phi})_2\end{bmatrix},$$

and further we get

$$\begin{aligned}\ddot{\xi}_1+a_{11}\alpha\dot{\xi}_1+a_{11}\beta\xi_1=&-a_{12}(\alpha\dot{\xi}_2+\beta\xi_2+(\dot{\mathcal{J}}\dot{\phi})_2)\\ &+(1-a_{11})(\dot{\mathcal{J}}\dot{\phi})_1\end{aligned}$$

and

$$\begin{aligned}\ddot{\xi}_2+a_{22}\alpha\dot{\xi}_2+a_{22}\beta\xi_2=&-a_{21}(\alpha\dot{\xi}_1+\beta\xi_1+(\dot{\mathcal{J}}\dot{\phi})_1)\\ &+(1-a_{22})(\dot{\mathcal{J}}\dot{\phi})_2,\end{aligned}$$

which can be regarded as a perturbed second-order constant coefficient differential equation with respect to ξ. In conclusion, pixel error ξ is able to converge exponentially. To illustrate the steady state of the system (Equation 14), further derivations continue to be given. As the pixel error decreases, all joint properties return to the interior of joint constraints. In this sense, joint properties, i.e., $\ddot{\phi}$, $\dot{\phi}$ and ϕ, are all inside the joint limits with $h_i=1$. Therefore, Equation (14) can be reorganized as

$$\ddot{\xi}(t)+\alpha\dot{\xi}(t)+\beta\xi(t)=0. \qquad (15)$$

It is worth mentioning that Equation (15) can be regarded as a second-order constant coefficient differential equation with regard to ξ. Moreover, the solutions of Equation (15) can be segmented into three subcases on account of different settings of α and β, given the original state $\xi(0) = \mathbf{p} - \mathbf{p}_{\mathrm{d}}$.

Subcase I: As for $\alpha^2 - 4\beta > 0$, the characteristic roots could be obtained simply as $\mathcal{R}_1 = (-\alpha + \sqrt{\alpha^2 - 4\beta})/2$ and $\mathcal{R}_2 = (-\alpha - \sqrt{\alpha^2 - 4\beta})/2$ with real number $\mathcal{R}_1 \neq \mathcal{R}_2$. Therefore, one can readily deduce

$$\xi(t) = \xi(0)(D_1 \exp(\mathcal{R}_1 t) + D_2 \exp(\mathcal{R}_2 t)),$$

with $D_1 = \alpha/(2\sqrt{\alpha^2 - 4\beta}) + 1/2$ and $D_2 = 1/2 - \alpha/(2\sqrt{\alpha^2 - 4\beta})$

Subcase II: As to $\alpha^2 - 4\beta = 0$, calculating characteristic roots generates $\mathcal{R}_1 = \mathcal{R}_2 = -\alpha/2$. Hence, it can be readily obtained:

$$\xi(t) = \xi(0)\exp(-\alpha/2t)(1 + \alpha/2t).$$

Subcase III: As to $\alpha^2 - 4\beta < 0$, we get two complex number roots as $\mathcal{R}_1 = \zeta + i\eta$ and $\mathcal{R}_2 = \zeta - i\eta$. Accordingly, it is evident that

$$\xi(t) = \xi(0)\exp(-\zeta t)(\cos(\eta t) - \zeta \sin(\eta t)/\eta).$$

The above three subcases indicate that the pixel error $\xi = \mathbf{p} - \mathbf{p}_{\mathrm{d}}$ converges to zero over time globally. The proof is complete.

4 DAIVS solution

The existing IBVS schemes, including the AIVS scheme (9), often require a detailed knowledge of the robot visual servoing system. However, in a non-ideal state, many unknown cases often exist, which can disturb the precise control of the robot, thus resulting in large errors. Recalling the data-driven learning algorithm (Equation 8), virtual visual Jacobian matrix $\bar{\mathcal{J}}$ is exploited to solve this issue.

4.1 DAIVS scheme and RNN solution

Based on the virtual visual Jacobian matrix, a DAIVS scheme (8) would be designed as

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\ddot{\phi}^{\mathrm{T}}\ddot{\phi} \\ \text{subject to} \quad & \ddot{\mathbf{p}} = \bar{\mathcal{J}}\ddot{\phi} + \dot{\bar{\mathcal{J}}}\dot{\phi} \\ & \mathbf{p} = \mathbf{p}_{\mathrm{d}} \\ & \ddot{\phi} \in \gamma. \end{aligned}$$

It is a remarkable fact that the DAIVS scheme does not involve the visual structure of the real robot. Instead, the virtual visual Jacobian matrix $\bar{\mathcal{J}}$ conveys the transformation relationship between the joint space and image space to deal with possible unknowns in the structure of the robot system. Compared with acceleration-level visual servoing schemes (Keshmiri et al., 2014; Anwar et al., 2019), the proposed DAIVS scheme offers two distinct advantages. First, it prioritizes the safety aspect by considering joint limits. Second, the DAIVS scheme takes into account the uncertainty of the robot vision system and employs the virtual visual Jacobian matrix for robot control, enhancing the fault tolerance ability. The existing acceleration-level visual servoing schemes (Keshmiri et al., 2014; Anwar et al., 2019) cannot accurately implement visual servoing tasks when the Jacobian matrix lacks precision. Furthermore, combining Equations (8, 13) generates

$$\ddot{\phi} = \mathcal{P}_{\gamma}(\bar{\mathcal{J}}^{\dagger}(-\alpha\dot{\mathbf{p}} - \beta(\mathbf{p} - \mathbf{p}_{\mathrm{d}}) - \dot{\bar{\mathcal{J}}}\dot{\phi})) \tag{16a}$$

$$\dot{\bar{\mathcal{J}}} = -\delta(\bar{\mathcal{J}}\dot{\phi} - \dot{\mathbf{p}})\dot{\phi}^{\mathrm{T}}. \tag{16b}$$

It is worth pointing out that the RNN (16) is divided into the inner cycle and outer cycle, i.e., the learning cycle and control cycle. Subsystem (Equation 16a), which can be viewed as the outer cycle, mainly generates the control signal to adjust the joint properties via virtual visual Jacobian matrix $\bar{\mathcal{J}}$. In return, inner cycle (Equation 16b) with learning ability can explore the relationship between end-effector motion and joint motion, thus producing virtual visual Jacobian matrix $\bar{\mathcal{J}}$ to simulate the movement process of real robots. From a control point of view, the inner cycle (Equation 16b) must converge faster than the outer cycle (Equation 16a). In this sense, $\delta \gg \alpha$ is a necessary condition for the normal operation of the system.

Note that both RNN (13) and RNN (16) involve the use of pseudo-inverse operations. As a result, various existing methods can be employed to mitigate singularity issues, such as the damped least squares method. Specifically, $\mathcal{J}^{\dagger}$ can be calculated via $\mathcal{J}^{\dagger} = \mathcal{J}^{\mathrm{T}}(\mathcal{J}\mathcal{J}^{\mathrm{T}} + h\mathcal{I})^{-1}$ with h being a tiny constant and $\mathcal{I}$ being an identity matrix. The additional item $h\mathcal{I}$ ensures that all eigenvalues of $\mathcal{J}\mathcal{J}^{\mathrm{T}} + h\mathcal{I}$ are never zero during the inversion process, thereby preventing singular issues. In addition, RNN (16) relies on the virtual visual Jacobian matrix and estimates the real Jacobian matrix using Equation (16b). This enables a robust handling of the visual system's uncertainty. However, RNN (13) relies on the real visual Jacobian matrix, leading to potential inaccuracies in the robot control process.

4.2 Stability analyses of RNN

The learning and control performance of the proposed DAIVS scheme aided with RNN (16) are proved by the following theorem.

Theorem 2: The Jacobian matrix error $E = \bar{\mathcal{J}} - \mathcal{J}$ and pixel error $\xi = \mathbf{p} - \mathbf{p}_{\mathrm{d}}$ produced by RNN (16) converges to zero, given a large enough δ.

Proof: The proof is segmented into two parts: (1) proving learning convergence; (2) proving control convergence.

Part 1: Proving learning convergence. Design the i-th system of Jacobian matrix error as $E_i = \bar{\mathcal{J}}_i - \mathcal{J}_i$ $(i = 1, 2)$ where $\bar{\mathcal{J}}_i$ and $\mathcal{J}_i$ denote the i-th row of $\bar{\mathcal{J}}$ and $\mathcal{J}$ and set the Lyapunov candidate $\mathcal{V}_i = (\bar{\mathcal{J}}_i - \mathcal{J}_i)(\bar{\mathcal{J}}_i - \mathcal{J}_i)^{\mathrm{T}}$. Calculating the time derivative of $\mathcal{V}_i$ leads to

$$\begin{aligned}\dot{\mathcal{V}}_i &= (\dot{\bar{\mathcal{J}}}_i - \dot{\mathcal{J}}_i)(\bar{\mathcal{J}}_i - \mathcal{J}_i)^{\mathrm{T}}\\ &= -\delta(\bar{\mathcal{J}}_i\dot{\phi} - \dot{\mathbf{p}}_i)\dot{\phi}^{\mathrm{T}}(\bar{\mathcal{J}}_i - \mathcal{J}_i)^{\mathrm{T}} - \dot{\mathcal{J}}_i(\bar{\mathcal{J}}_i - \mathcal{J}_i)^{\mathrm{T}}\\ &= -\delta(\bar{\mathcal{J}}_i\dot{\phi} - \mathcal{J}_i\dot{\phi})\dot{\phi}^{\mathrm{T}}(\bar{\mathcal{J}}_i - \mathcal{J}_i)^{\mathrm{T}} - \dot{\mathcal{J}}_i(\bar{\mathcal{J}}_i - \mathcal{J}_i)^{\mathrm{T}}\\ &\leq -\delta\Pi(\dot{\phi}\dot{\phi}^{\mathrm{T}})(\bar{\mathcal{J}}_i - \mathcal{J}_i)(\bar{\mathcal{J}}_i - \mathcal{J}_i)^{\mathrm{T}} - \dot{\mathcal{J}}_i(\bar{\mathcal{J}}_i - \mathcal{J}_i)^{\mathrm{T}},\end{aligned}$$

where $\dot{\mathbf{p}}_i$ represents the i-th element of $\dot{\mathbf{p}}$, and $\Pi(\dot{\phi}\dot{\phi}^{\mathrm{T}})$ denotes the least eigenvalue of matrix $\dot{\phi}\dot{\phi}^{\mathrm{T}}$. When the manipulator is tracking the feature, the value of $\Pi(\dot{\phi}\dot{\phi}^{\mathrm{T}})$ is always greater than zero. In this case, we substitute $E_i = \bar{\mathcal{J}}_i - \mathcal{J}_i$ into the above equation, resulting in the following expression:

$$\begin{aligned}\dot{\mathcal{V}}_i &\leq -\delta\Pi(\dot{\phi}\dot{\phi}^{\mathrm{T}})E_iE_i^{\mathrm{T}} - \dot{\mathcal{J}}_iE_i^{\mathrm{T}}\\ &\leq -\delta\Pi(\dot{\phi}\dot{\phi}^{\mathrm{T}})||E_i||_2^2 + ||\dot{\mathcal{J}}_i||_2||E_i||_2\\ &= ||E_i||_2(||\dot{\mathcal{J}}_i||_2 - \delta\Pi(\dot{\phi}\dot{\phi}^{\mathrm{T}})||E_i||_2).\end{aligned}$$

For further analysis, we consider three cases based on the above equation:

- If $||E_i||_2 > ||\dot{\mathcal{J}}_i||_2/\delta\Pi(\dot{\phi}\dot{\phi}^{\mathrm{T}})$, we observe $\dot{\mathcal{V}}_i < 0$ and $\mathcal{V}_i > 0$. This indicates that in this case, E_i converges until $||E_i||_2 = ||\dot{\mathcal{J}}_i||_2/\delta\Pi(\dot{\phi}\dot{\phi}^{\mathrm{T}})$.
- If $||E_i||_2 = ||\dot{\mathcal{J}}_i||_2/\delta\Pi(\dot{\phi}\dot{\phi}^{\mathrm{T}})$, we find $\dot{\mathcal{V}}_i \leq 0$ and $\mathcal{V}_i > 0$. This implies that E_i will continue to converge or remain at the state with $||E_i||_2 = ||\dot{\mathcal{J}}_i||_2/\delta\Pi(\dot{\phi}\dot{\phi}^{\mathrm{T}})$.
- If $||E_i||_2 < ||\dot{\mathcal{J}}_i||_2/\delta\Pi(\dot{\phi}\dot{\phi}^{\mathrm{T}})$, we have two possibilities: either $\dot{\mathcal{V}}_i > 0$ and $\mathcal{V}_i > 0$, or $\dot{\mathcal{V}}_i \leq 0$ and $\mathcal{V}_i > 0$. In the former possibility, the error will increase until $||E_i||_2= ||\dot{\mathcal{J}}_i||_2/\delta\Pi(\dot{\phi}\dot{\phi}^{\mathrm{T}})$. In the latter possibility, the error will continue to converge or remain constant.

Combining the above three cases, it can be summarized that $\lim_{t\to+\infty}||E_i||_2 \leq ||\dot{\mathcal{J}}_i||_2/\delta\Pi(\dot{\phi}\dot{\phi}^{\mathrm{T}})$. Furthermore, it can be deduced that the Jacobian matrix error $E = \bar{\mathcal{J}} - \mathcal{J}$ produced by RNN (16a) globally approach zero, given a sufficiently large value of δ.

Part 2: Proving control convergence.

According to the proof in *Part 1*, we take advantage of the LaSalle's invariant principle (Khalil, 2001) again to conduct the convergence proof on Equation (16b). In other words, the following formula is provided by replacing $\bar{\mathcal{J}}$ and $\dot{\bar{\mathcal{J}}}$ with $\mathcal{J}$ and $\dot{\mathcal{J}}$:

$$\ddot{\phi} = \mathcal{P}_{\gamma}(\mathcal{J}^{\dagger}(-\alpha\dot{\mathbf{p}} - \beta(\mathbf{p} - \mathbf{p}_{\mathrm{d}}) - \dot{\mathcal{J}}\dot{\phi})), \tag{17}$$

which is equivalent to Equation (13). In consequence, the proof on the convergence of the pixel error $\mathbf{p} - \mathbf{p}_{\mathrm{d}}$ in Equation (17) has been discussed in Theorem 1 and thus omitted here. The proof is complete.

5 Simulation verifications

In this section, simulations are conducted on a Franka Emika Panda manipulator with 7 DOFs for completing a visual servoing task, which are synthesized by the proposed AIVS scheme (9) and the proposed DAIVS scheme. Note that the AIVS scheme (9) is able to drive the redundant manipulator to perform the visual servoing task with a given visual Jacobian matrix, and that, the DAIVS scheme can deal with the unknown situation in the robot system dynamically in the absence of the visual Jacobian matrix. For the simulations, this study utilizes a computer with an Intel Core i7-12700 processor and 32 GB RAM. The simulations are performed using MATLAB/Simulink software version R2022a.

First, some necessary information and parameter settings about the manipulator and camera structure are given below. The Franka Emika Panda manipulator is a 7-DOF redundant manipulator (Gaz et al., 2019), with a camera mounted on its end-effector. In addition, we set $l = 8 \times 10^{-3}$ m, $u_0 = v_0 = 256$ pixel, $a_x = a_y = 8 \times 10^4$ pixel/m, and design $\mu = \nu = 20$ with $z = 2$, task execution time $T = 20$ s and $\mathbf{p}_{\mathrm{d}} = [256, 256]^{\mathrm{T}}$ pixel. In addition, the joint limits are set as $\ddot{\phi}^{+} = -\ddot{\phi}^{-} = [2]_{7\times1}$ rad/s^2, $\dot{\phi}^{+} = -\dot{\phi}^{-} = [0.6]_{7\times1}$ rad/s, $\phi^{+} = -\phi^{-} = [2.5]_{7\times1}$ rad and $\theta = [0.076]_{7\times1}$ rad. It is noteworthy that the parameters can be divided into two categories: structural parameters and convergence parameters. Structural parameters, such as l, u_0, v_0, a_x, and a_y, are dependent on the configuration of the visual servo system. On the other hand, the convergence parameters, namely, μ, ν, α, β, and δ, play a vital role in adjusting the convergence behavior of RNN (16). These convergence parameters are set to values greater than zero, and their specific values can be determined through the trial and error method.

5.1 Simulation of AIVS scheme

In this subsection, in order to prove the feasibility of the AIVS scheme (9), four simulations with different initial position states of the Franka Emika Panda manipulator are conducted to trace one desired feature with results shown in Figure 3. Simply design $\alpha = 10$ and $\beta = 10$. It would be readily discovered from Figure 3A that four test examples from four different directions are straightforward to successfully pursue the desired pixel. With test 4 as an example, detailed joint data and pixel errors are shown in Figure 3B through Figure 3F, which illustrate that the joint angle, joint velocity, and joint acceleration are all kept inside the joint limit and that the pixel error can converge to zero within 5 s. The above descriptions well verify the validity of the proposed AIVS scheme (9) in the case of the known visual servoing Jacobian matrix to solve the visual servoing problem at the acceleration level.

5.2 Simulation of DAIVS scheme

This subsection indicates the feasibility and capability of the pixel error convergence of the DAIVS scheme aided with the RNN (16) by providing simulation results, as shown in Figure 4. Furthermore, we choose $\delta = 2 \times 10^4$, $\alpha = 10$ and $\beta = 40$. Notably, the virtual visual Jacobian matrix is exploited with random initial values, instead of the real visual Jacobian matrix to facilitate system operation. The end-effector of the robotic arm is oriented toward the object, as shown in Figure 4A. In addition, the joint acceleration is shown in Figure 4B, which is confined to the joint limit and maintain the normal operation. As shown in Figure 4C, the Franka

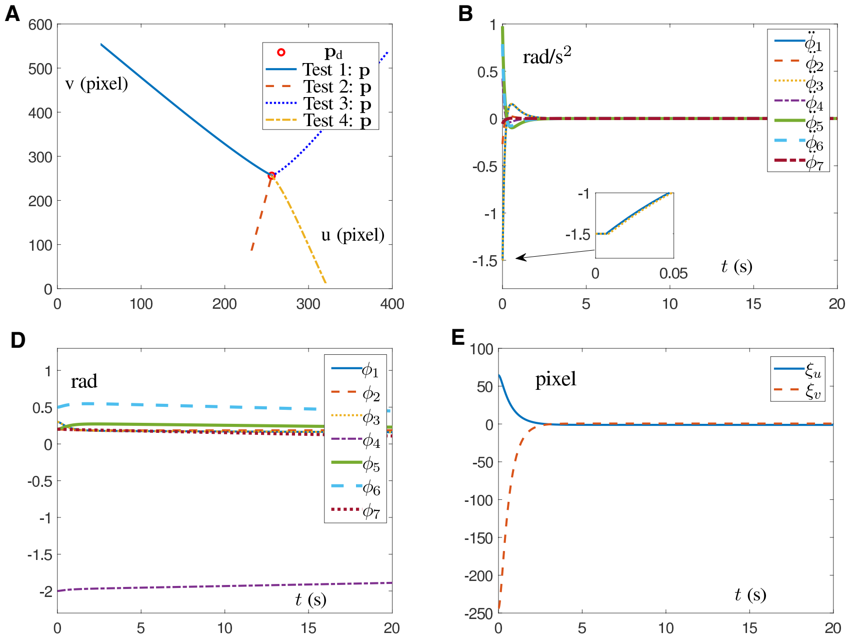


FIGURE 3
Simulations on a Franka Emika Panda manipulator carrying out IBVS task synthesized by the AIVS scheme (9) assisted by RNN (13) with four test examples. **(A)** Profiles of feature trajectories and desired pixel point in four tests. **(B)** Profiles of joint acceleration in test 4. **(C)** Profiles of joint velocity in test 4. **(D)** Profiles of joint angle in test 4. **(E)** Profiles of pixel error in test 4. **(F)** Profiles of Euclidean norm of pixel error in test 4.

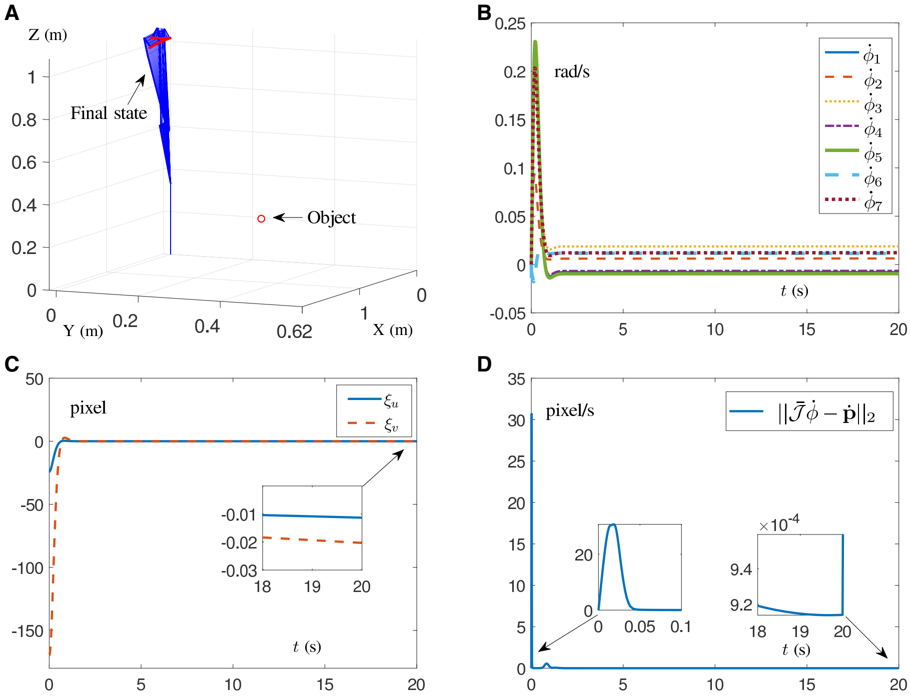


FIGURE 4
Simulation results on a Franka Emika Panda manipulator with unknown structure carrying out the IBVS task synthesized by the DAIVS scheme assisted with RNN (16). **(A)** Profiles of the movement process. **(B)** Profiles of the joint acceleration. **(C)** Profiles of the pixel error. **(D)** Profiles of the Euclidean norm of learning error.

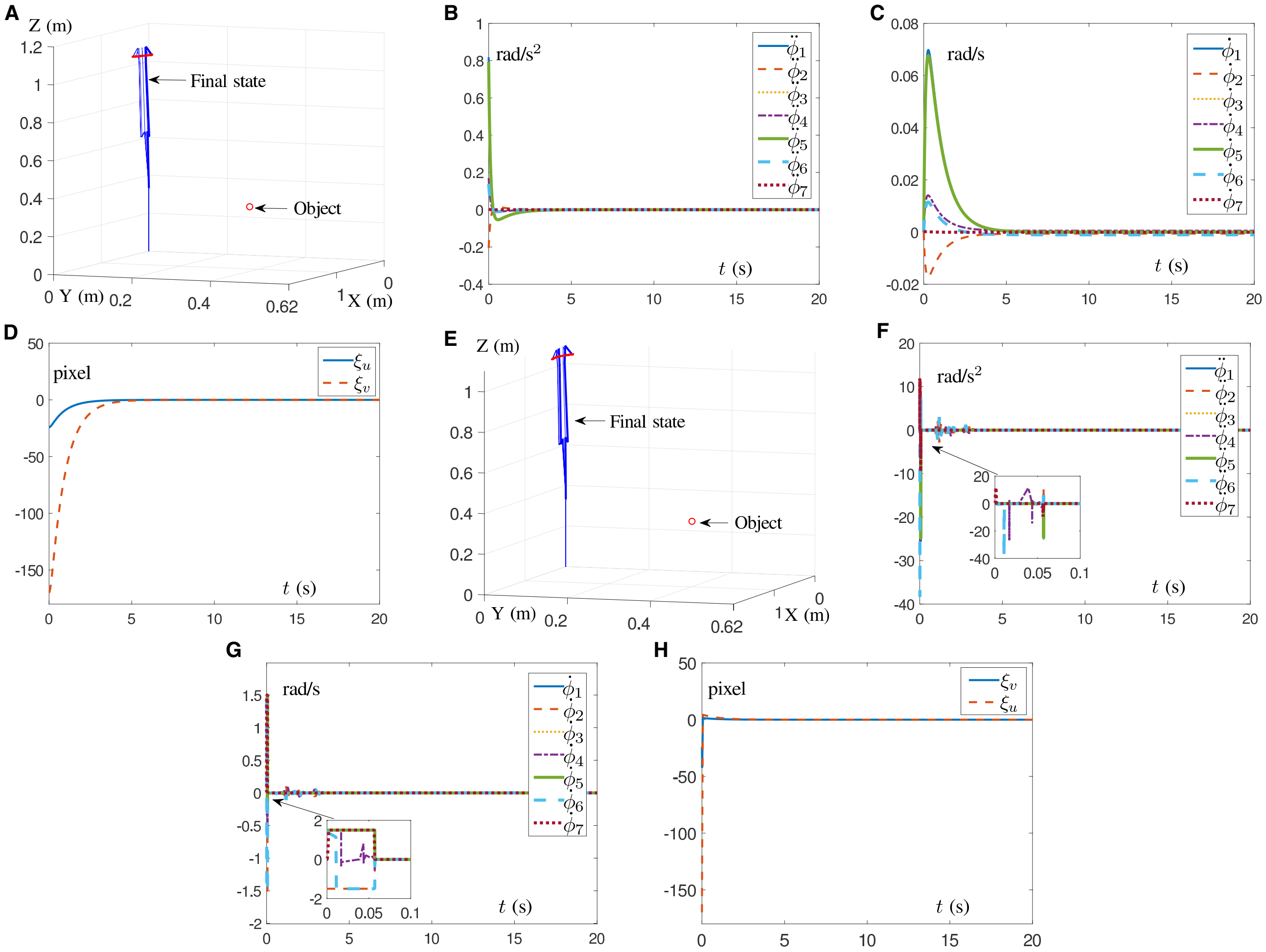


FIGURE 5
Simulation results on a Franka Emika Panda manipulator with accurate structure information carrying out IBVS task. **(A)** Profiles of motion process assisted with RNN (13). **(B)** Profiles of joint acceleration assisted with RNN (13). **(C)** Profiles of joint velocity assisted with RNN (13). **(D)** Profiles of pixel error assisted with RNN (13). **(E)** Profiles of motion process assisted with RNN (18). **(F)** Profiles of joint acceleration assisted with RNN (18). **(G)** Profiles of joint velocity assisted with RNN (18). **(H)** Profiles of pixel error assisted with RNN (18).

Emika Panda manipulator successfully traces the desired feature with pixel error converging to zero and maintaining the order of 10^{-2} pixel. As for the learning ability, Figure 4D illustrates that the virtual robot manipulator can learn the movement of the real robot manipulator with the learning error approaching to zero in 0.05 s and maintaining the order of 10^{-4} pixel/s. In short, the simulation results in Figure 4 highlight the simultaneous learning and control ability of RNN (16).

5.3 Comparisons of proposed schemes

This subsection offers simulation comparison results between the proposed schemes aided with the corresponding RNNs and the IBVS method presented in the study by Zhang and Li (2018). In this regard, the RNN provided in the study by Zhang and Li (2018) is shown as

$$\dot{\phi} = \mathcal{P}_{\gamma}(-\kappa_1\mathcal{J}^{\mathrm{T}}(\mathbf{p}-\mathbf{p}_{\mathrm{d}}) - \kappa_2\mathcal{J}^{\mathrm{T}}\int_0^t(\mathbf{p}-\mathbf{p}_{\mathrm{d}})\mathrm{d}t), \qquad (18)$$

where parameters $\kappa_1 > 0$ and $\kappa_2 > 0$ determine the rate of error convergence. It is worth pointing out that the IBVS method in the study by Zhang and Li (2018) assisted with RNN (18) is constructed from the viewpoint of the velocity level, and that, RNN (18) requires exact structural information $\mathcal{J}$ to maintain the normal operation.

In the first place, simulations are conducted on the Franka Emika Panda manipulator for IBVS task with Figures 5A–D synthesized by RNN (13) and Figures 5E–H synthesized by RNN (18). Notably, the results in Figure 5 are carried out on the premise of known structural information $\mathcal{J}$ with parameters $\kappa_1 = \kappa_2 = 2$, $\alpha = 10$, and $\beta = 10$. As shown in Figures 5A, E, the manipulator's end-effector is controlled to point toward the object. In Figure 5B, the joint acceleration generated by RNN (13) is safely confined within the joint limits, while the joint acceleration generated by RNN (18) exists a sudden change of ∼38 rad/s² in Figure 5F, which may cause damage to the robot. Furthermore, in contrast to Figure 5G, the joint velocity shown in Figure 5C is smaller and exhibits smoother changes, making it more suitable for real-world scenarios. Figures 5D, H demonstrate that both RNN (13) and RNN

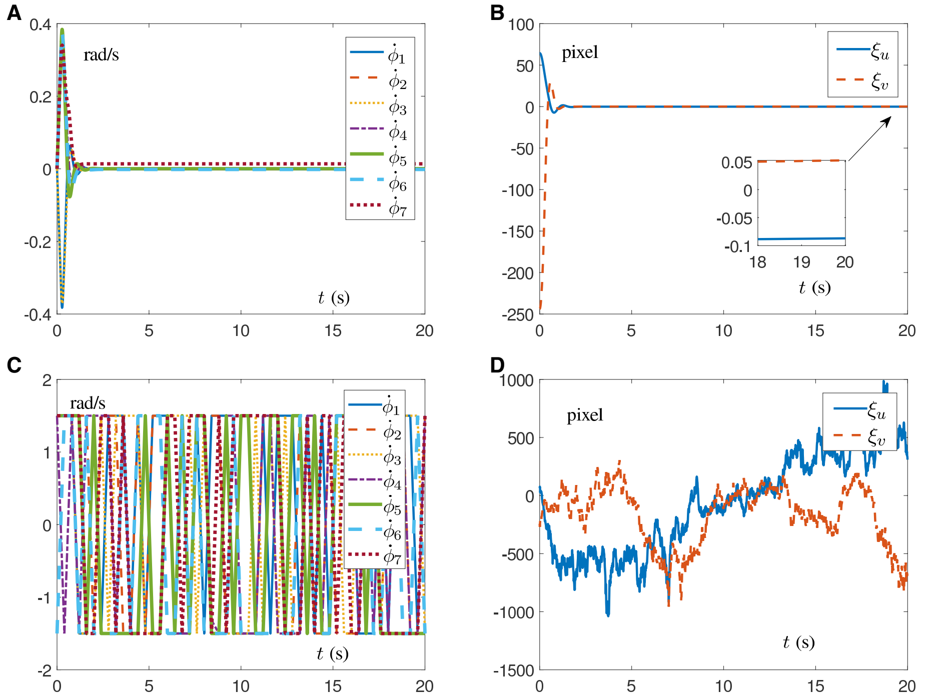


FIGURE 6
Simulation results on a Franka Emika Panda manipulator carrying out IBVS task with unknown structure. **(A)** Profiles of joint velocity assisted with RNN (16). **(B)** Profiles of pixel error assisted with RNN (16). **(C)** Profiles of joint velocity assisted with RNN (18). **(D)** Profiles of pixel error assisted with RNN (18).

TABLE 1 Comparisons among different approaches for visual servoing of robot manipulators.

	Visual servoing	Scheme level	Velocity constraints	Acceleration constraints	Structure information	Jacobian matrix learning
RNN (13)	Yes	Acceleration	Yes	Yes	Unnecessary	Yes
RNN (16)	Yes	Acceleration	Yes	Yes	Necessary	No
Van et al. (2018)	Yes	Velocity	No	No	Necessary	No
Hashimoto et al. (1991)	Yes	Velocity	No	No	Necessary	No
Zhang et al. (2017)	Yes	Velocity	Yes	No	Necessary	No
Zhang and Li (2018)	Yes	Velocity	Yes	No	Necessary	No
Li et al. (2020)	Yes	Velocity	Yes	No	Necessary	No
Keshmiri et al. (2014)	Yes	Acceleration	No	No	Necessary	No
Anwar et al. (2019)	Yes	Acceleration	No	No	Necessary	No
Zhu et al. (2022)	Yes	Torque	No	No	Necessary	No

(18) are able to quickly propel pixel errors to zero. Therefore, it is concluded from the above results that AIVS scheme (9) aided by RNN (13) is able to guarantee a better safety performance when controlling the manipulator.

Beyond that, in the case of the unknown visual system, corresponding comparison simulations are driven by the DAIVS scheme aided with the RNN (16) and the IBVS method in the study by Zhang and Li (2018) assisted with RNN (18). The results are shown in Figure 6 with parameters $\kappa_1 = \kappa_2 = 2$, $\alpha = 10$, $\beta = 40$, and $\delta = 2 \times 10^4$. To simulate the unknown visual system, $\bar{\mathcal{J}}$ in Equation (16) and $\mathcal{J}$ in Equation (18) are random matrices of constants with the absolute value of each element < 100. Figures 6A, B well embody that, when encountering unknown structural information, the DAIVS scheme assisted with RNN (16) controls the Franka Emika Panda manipulator to preferably complete IBVS task with the pixel error converging to zero. Nevertheless, the generated joint velocity in Figure 6C changes dramatically within the joint limit in a mess. Even worse, the

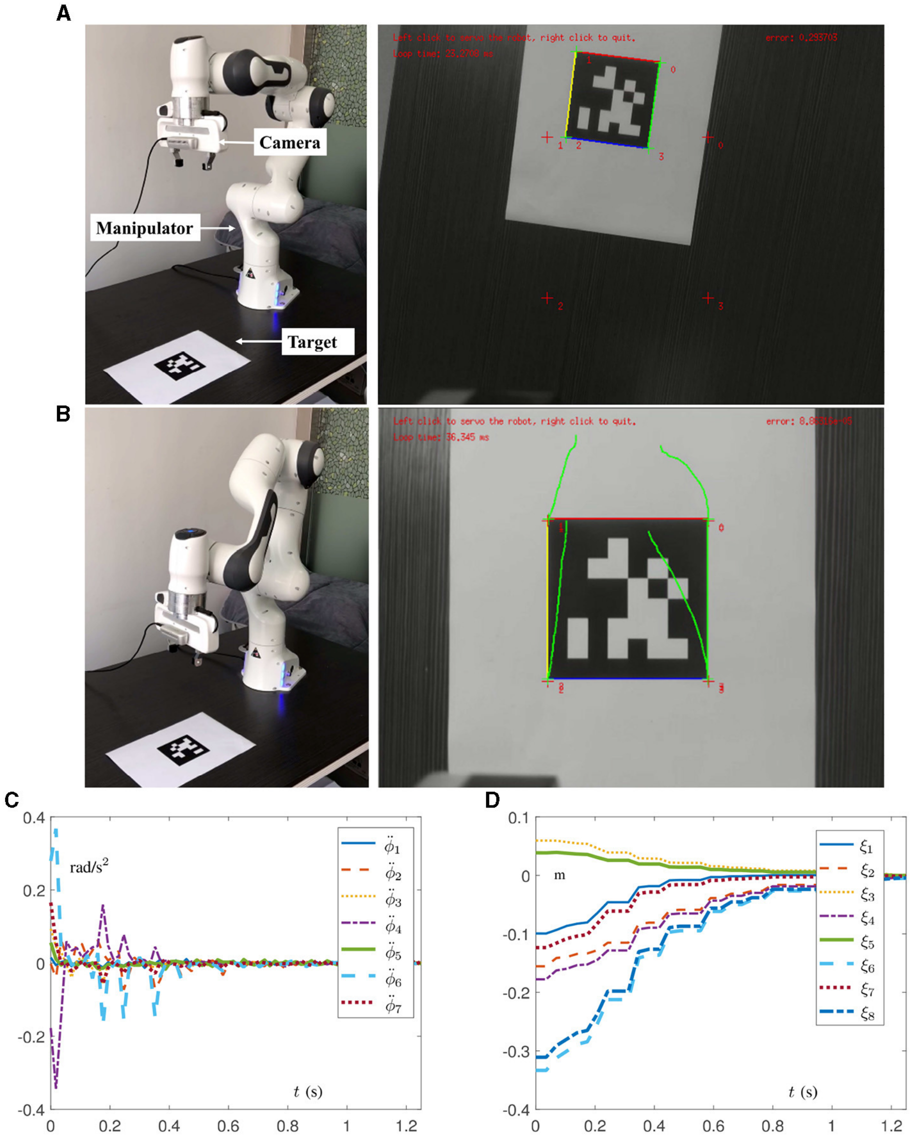


FIGURE 7
Physical experiments on a Franka Emika Panda manipulator assisted with RNN (16) for carrying out IBVS task with a fixed target. **(A)** Initial states of the manipulator and camera. **(B)** Final states of the manipulator and camera. **(C)** Profiles of joint acceleration. **(D)** Profiles of tracking error.

pixel error driven by RNN (18) does not converge and maintain a diffused state, which indicates the failure of the IBVS task. In conclusion, the proposed DAIVS scheme is able to deal with the unknown structural information in the robot system and fulfill the visual servo control with simultaneous learning and control performance.

Furthermore, comparison results among different existing approaches (Hashimoto et al., 1991; Keshmiri et al., 2014; Zhang et al., 2017; Van et al., 2018; Zhang and Li, 2018; Anwar et al., 2019; Li et al., 2020; Zhu et al., 2022) for visual servoing of robot manipulators are presented in Table 1. It is worth emphasizing that, compared with the prior art, the proposed RNN (13) and RNN (16) are the first acceleration-level work, considering the multiple levels of joint constraints, and RNN (16) is the first study to dispose the unknown situations in the robot visual system with simultaneous learning and control ability. As a result, the above two points are the innovative contributions of this study.

6 Experiments on real manipulators

To verify the effectiveness and practicability of the proposed DAIVS scheme, physical experiments on a real manipulator are conducted in this section, which are driven by the DAIVS scheme aided with RNN (16). Specifically, the experiments essentially rely on C++ and the visual servoing platform (ViSP) for embedding algorithms and control (Marchand et al., 2005), which are built on ubuntu 16.04 LTS operating system. In addition, the experiment platform consists of a Franka Emika Panda manipulator, an Intel RealSense Camera D435i, a personal computer, and an AprilTag (target). It is worth mentioning that the acceleration control

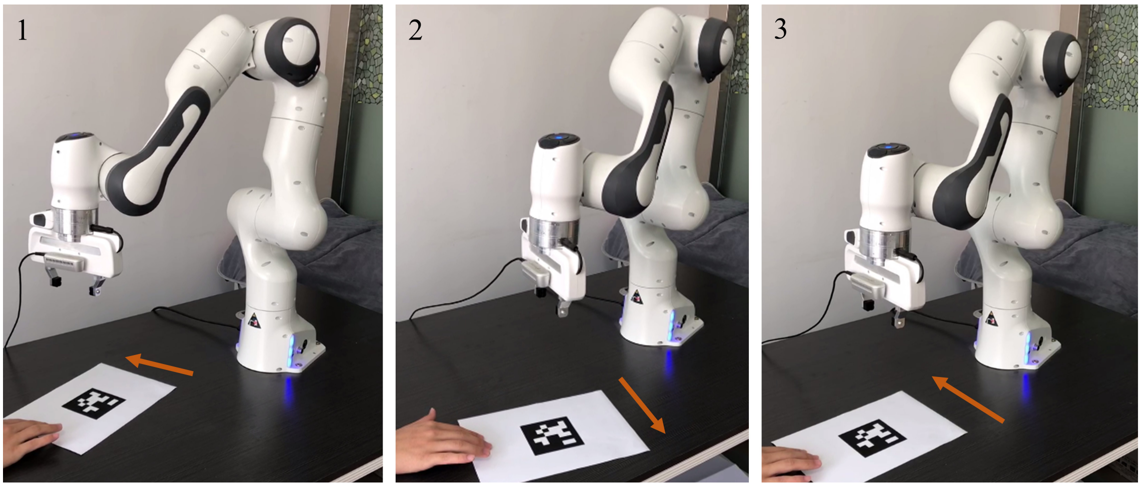


FIGURE 8
Physical experiments on a Franka Emika Panda manipulator assisted with RNN (16) for carrying out IBVS task with a moving target.

commands generated by the proposed RNN (16) are transmitted in a discrete form with a frequency of 1,000 Hz, and parameter settings of RNN (16) are designed as follows. We choose $\alpha = 10$, $\beta = 10$, $\delta = 10^6$, $\mu = \nu = 20$, $\ddot{\phi}^+ = -\ddot{\phi}^- = [15,\ 7.5,\ 10,\ 12.5,\ 15,\ 20,\ 20]^{\mathrm{T}}$ rad/s^2, $\dot{\phi}^+ = -\dot{\phi}^- = [2.1, 2.1, 2.1, 2.1, 2.6, 2.6, 2.6]^{\mathrm{T}}$ rad/s, $\phi^+ = [2.8,\ 1.7,\ 2.8,\ -0.1,\ 2.9,\ 3.7,\ 2.8]^{\mathrm{T}}$ rad, $\phi^- = [-2.8,\ -1.7,\ -2.8,\ -3.0,\ -2.8,\ -0,\ -2.8]^{\mathrm{T}}$ rad, and $\bar{\mathcal{J}}(0) = \mathcal{J}(0)$. As for the parameter settings of the camera and pixel coordinates, they can be directly referenced to ViSP (Marchand et al., 2005). Different from the previous simulations, the physics experiments set the target as an AprilTag containing four features. As a result, the physical parameters associated with the features are expanded to 8 instead of 2.

Experiment results on the Franka Emika Panda manipulator tracking the fixed target are shown in Figures 7, 8 with $\mathbf{p}_{\mathrm{d}} = [-0.06,\ -0.06,\ 0.06,\ -0.06,\ 0.06,\ 0.06,\ -0.06,\ 0.06]^{\mathrm{T}}$ m for the given task in the camera system. It is worth mentioning that the robot manipulator adjusts the joint state to recognize and approach the target, and when the pixel error reaches the order of 10^{-5} pixel, the task automatically completes. It is important that the whole process of learning and control does not involve the real Jacobian matrix to simulate the situation of the unknown structure. In Figures 7A, B, the initial and final states of the manipulator and camera indicate that the visual servoing task is successfully realized by the DAIVS scheme with execution time of 1.25 s. Specifically, the joint acceleration in Figure 7C varies normally within the joint constraints. In the meantime, the tracking errors ξ of four features are presented in Figure 7D, which illustrate the precise control ability of the DAIVS scheme with global convergence to zero.

Beyond that, experiments on the Franka Emika Panda manipulator tracking the moving target are conducted to demonstrate the feasibility of the DAIVS scheme. In Figure 8, the AprilTag is moved artificially by the hand toward the left and right and simultaneously the manipulator constantly adjusts joint states to achieve the characteristics of real-time visual tracking. More vividly, the experiment videos corresponding to Figures 7, 8 are available at https://youtu.be/6uw35bidVcw.

7 Conclusion

This study has proposed an AIVS scheme for robot manipulators, taking into account joint limits at multiple levels. On this basis, incorporating data-driven techniques, a DAIVS scheme has been proposed to handle potential unknown situations in the robot visual system. Furthermore, RNNs have been exploited to generate the online solution corresponding to the proposed schemes with theoretical analyses, demonstrating the simultaneous learning and control ability of the proposed DAIVS scheme. Then, numerous simulations and experiments have been carried out on a Franka Emika Panda manipulator to track the desired feature. The results validate the theoretical analyses, demonstrate the feasibility of the AIVS scheme, and showcase the fast convergence and robustness of the DAIVS scheme. Compared with the method in the study by Zhang and Li (2018), the DAIVS scheme exhibits superior learning capability and achieves visual servoing control with the unknown Jacobian matrix.

In summary, this study provides a data-driven approach for the precise manipulation of robots in IBVS tasks, addressing unknown situations that could affect the robot's Jacobian matrix. In the future, we aim to expand our research to incorporate dynamic factors, utilizing joint torque as control signals and considering dynamic uncertainties.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

LW: Investigation, Methodology, Resources, Software, Writing – original draft. ZX: Data curation, Investigation, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Key Laboratory of Tibetan Information Processing, Ministry of Education, under Grant 2024-Z-002, and supported by the Supercomputing Center of Lanzhou University.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Agarwal, S., and Akella, S. (2024). Line coverage with multiple robots: algorithms and experiments. *IEEE Transact. Robot.* 40, 1664–1683. doi: 10.1109/TRO.2024.3355802

Anwar, A., Lin, W., Deng, X., Qiu, J., and Gao, H. (2019). Quality inspection of remote radio units using depth-free image-based visual servo with acceleration command. *IEEE Transact. Ind. Electron.* 66, 8214–8223. doi: 10.1109/TIE.2018.2881948

Cigliano, P., Lippiello, V., Ruggiero, F., and Siciliano, B. (2015). Robotic ball catching with an eye-in-hand single-camera system. *IEEE Transact. Cont. Syst. Technol.* 23, 1657–1671. doi: 10.1109/TCST.2014.2380175

Gaz, C., Cognetti, M., Oliva, A., Robuffo Giordano, P., and De Luca, A. (2019). Dynamic identification of the Franka Emika Panda robot with retrieval of feasible parameters using penalty-based optimization. *IEEE Robot. Automat. Lett.* 4, 4147–4154. doi: 10.1109/LRA.2019.2931248

Hashimoto, K., Kimoto, T., Ebine, T., and Kimura, H. (1991). "Manipulator control with image-based visual servo," in *Proceedings IEEE International Conference on Robotics and Automation* (New York, NY: Sacramento, CA: IEEE), 2267–2271.

Jin, P., Lin, Y., Song, Y., Li, T., and Yang, W. (2023). Vision-force-fused curriculum learning for robotic contact-rich assembly tasks. *Front. Neurorobot.* 17:1280773. doi: 10.3389/fnbot.2023.1280773

Jin, L., Liu, L., Wang, X., Shang, M., and Wang, F.-Y. (2024). Physical-informed neural network for MPC-based trajectory tracking of vehicles with noise considered. *IEEE Transact. Intell. Vehicl.* doi: 10.1109/TIV.2024.3358229

Keshmiri, M., Xie, W.-F., and Mohebbi, A. (2014). Augmented image-based visual servoing of a manipulator using acceleration command. *IEEE Transact. Ind. Electron.* 61, 5444–5452. doi: 10.1109/TIE.2014.2300048

Khalil, H. K. (2001). *Nonlinear Systems*. Englewood Cliffs, NJ: Prentice Hall.

Li, S., Shao, Z., and Guan, Y. (2019). A dynamic neural network approach for efficient control of manipulators. *IEEE Transact. Syst. Man Cybernet. Syst.* 49, 932–941. doi: 10.1109/TSMC.2017.2690460

Li, W., Chiu, P. W. Y., and Li, Z. (2020). An accelerated finite-time convergent neural network for visual servoing of a flexible surgical endoscope with physical and RCM constraints. *IEEE Transact. Neural Netw. Learn. Syst.* 31, 5272–5284. doi: 10.1109/TNNLS.2020.2965553

Liang, X., Wang, H., Liu, Y.-H., Chen, W., and Jing, Z. (2018). Image-based position control of mobile robots with a completely unknown fixed camera. *IEEE Trans. Automat. Contr.* 63, 3016–3023. doi: 10.1109/TAC.2018.2793458

Liufu, Y., Jin, L., Shang, M., Wang, X., and Wang, F.-Y. (2024). ACP-incorporated perturbation-resistant neural dynamics controller for autonomous vehicles. *IEEE Transact. Intelli. Vehicl.* doi: 10.1109/TIV.2023.3348632

Malis, E., Mezouar, Y., and Rives, P. (2010). Robustness of image-based visual servoing with a calibrated camera in the presence of uncertainties in the three-dimensional structure. *IEEE Transact. Robot.* 26, 112–120. doi: 10.1109/TRO.2009.2033332

Marchand, E., Spindler, F., and Chaumette, F. (2005). ViSP for visual servoing: a generic software platform with a wide class of robot control skills. *IEEE Robot. Automat. Mag.* 12, 40–52. doi: 10.1109/MRA.2005.1577023

Na, J., Yang, J., Wang, S., Gao, G., and Yang, C. (2021). Unknown dynamics estimator-based output-feedback control for nonlinear pure-feedback systems *IEEE Transact. Syst. Man Cybernet. Syst.* 51, 3832–3843. doi: 10.1109/TSMC.2019.2931627

Park, D.-H., Kwon, J.-H., and Ha, I.-J. (2012). Novel position-based visual servoing approach to robust global stability under field-of-view constraint. *IEEE Transact. Ind. Electron.* 59, 4735–4752. doi: 10.1109/TIE.2011.2179270

Peng, G., Chen, C. L. P., and Yang, C. (2023). Robust admittance control of optimized robot-environment interaction using reference adaptation. *IEEE Transact. Neural Netw. Learn. Syst.* 34, 5804–5815. doi: 10.1109/TNNLS.2021.313 1261

Stanimirovic, P. S., Zivkovic, I. S., and Wei, Y. (2015). Recurrent neural network for computing the Drazin inverse. *IEEE Transact. Neural Netw. Learn. Syst.* 26, 2830–2843. doi: 10.1109/TNNLS.2015.2397551

Tang, Z., and Zhang, Y. (2022). Refined self-motion scheme with zero initial velocities and time-varying physical limits via Zhang neurodynamics equivalency. *Front. Neurorobot.* 16:945346. doi: 10.3389/fnbot.2022.945346

Van, M., Ge, S. S., and Ceglarek, D. (2018). Fault estimation and accommodation for virtual sensor bias fault in image-based visual servoing using particle filter. *IEEE Transact. Ind. Inf.* 14, 1312–1322. doi: 10.1109/TII.2017.2723930

Wang, X., Sun, Y., Xie, Y., Bin, J., and Xiao, J. (2023). Deep reinforcement learning-aided autonomous navigation with landmark generators. *Front. Neurorobot.* 17:1200214. doi: 10.3389/fnbot.2023.1200214

Xie, Z., Jin, L., Luo, X., Hu, B., and Li, S. (2022). An acceleration-level data-driven repetitive motion planning scheme for kinematic control of robots with unknown structure. *IEEE Transact. Syst. Man Cybernet. Syst.* 5152, 5679–5691. doi: 10.1109/TSMC.2021.3129794

Xie, Z., and Jin, L. (2023). A fuzzy neural controller for model-free control of redundant manipulators with unknown kinematic parameters. *IEEE Transact. Fuzzy Syst.* 32, 1589–1601. doi: 10.1109/TFUZZ.2023.3328545

Xu, C., Sun, Z., Wang, C., Wu, X., Li, B., and Zhao, L. (2023). An advanced bionic knee joint mechanism with neural network controller. *Front. Neurorobot.* 17:1178006. doi: 10.3389/fnbot.2023.1178006

Xu, F., Wang, H., Wang, J., Au, K. W. S., and Chen, W. (2019). Underwater dynamic visual servoing for a soft robot arm with online distortion correction. *IEEE ASME Transact. Mechatron.* 24, 979–989. doi: 10.1109/TMECH.2019.2908242

Yang, C., Chen, C., He, W., Cui, R., and Li, Z. (2019). Robot learning system based on adaptive neural control and dynamic movement primitives. *IEEE Transact. Neural Netw. Learn. Syst.* 30, 777–787. doi: 10.1109/TNNLS.2018.2852711

Zeng, D., Liu, Y., Qu, C., Cong, J., Hou, Y., and Lu, W. (2024). Design and human-robot coupling performance analysis of flexible ankle rehabilitation robot. *IEEE Robot. Automat. Lett.* 9, 579–586. doi: 10.1109/LRA.2023.3330052

Zeng, Z., Wang, J., and Liao, X. (2003). Global exponential stability of a general class of recurrent neural networks with time-varying delays. *IEEE Transact. Circ. Syst. I Fund. Theory Appl.* 50, 1353–1358. doi: 10.1109/TCSI.2003.817760

Zhang, Y., and Li, S. (2018). A neural controller for image-based visual servoing of manipulators with physical constraints. *IEEE Transact. Neural Netw. Learn. Syst.* 29, 5419–5429. doi: 10.1109/TNNLS.2018.2802650

Zhang, Y., and Li, S. (2023). Kinematic control of serial manipulators under false data injection attack. *IEEE CAA J. Automat. Sin.* 10, 1009–1019. doi: 10.1109/JAS.2023.123132

Zhang, Y., Li, S., Liao, B., Jin, L., and Zheng, L. (2017). "A recurrent neural network approach for visual servoing of manipulators," in *2017 IEEE International Conference on Information and Automation (ICIA)* (New York, NY: Macao: IEEE), 614–619.

Zhang, Z., and Zhang, Y. (2012). Acceleration-level cyclic-motion generation of constrained redundant robots tracking different paths. *IEEE Transact. Syst. Man Cybernet. Part B* 42, 1257–1269. doi: 10.1109/TSMCB.2012.2189003

Zheng, X., Liu, M., Jin, L., and Yang, C. (2024). Distributed collaborative control of redundant robots under weight-unbalanced directed graphs. *IEEE Transact. Ind. Inf.* 20, 681–690. doi: 10.1109/TII.2023.3268778

Zhu, M., Huang, C., Qiu, Z., Zheng, W., and Gong, D. (2022). Parallel image-based visual servoing/force control of a collaborative delta robot. *Front. Neurorobot.* 16:922704. doi: 10.3389/fnbot.2022.922704

frontiers | Frontiers in Neurorobotics

TYPE Review
PUBLISHED 05 April 2024
DOI 10.3389/fnbot.2024.1385778

Advancing autonomy through lifelong learning: a survey of autonomous intelligent systems

Dekang Zhu[1], Qianyi Bu[2], Zhongpan Zhu[1,3]*, Yujie Zhang[1] and Zhipeng Wang[1]

[1]College of Electronic and Information Engineering, Tongji University, Shanghai, China, [2]College of Science and Engineering, University of Glasgow, Glasgow, United Kingdom, [3]College of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai, China

OPEN ACCESS

EDITED BY
Long Jin,
Lanzhou University, China

REVIEWED BY
Bolin Liao,
Jishou University, China
Jingtai Li,
Jiaying University, China

*CORRESPONDENCE
Zhongpan Zhu
521bergsteiger@tongji.edu.cn

RECEIVED 13 February 2024
ACCEPTED 25 March 2024
PUBLISHED 05 April 2024

CITATION
Zhu D, Bu Q, Zhu Z, Zhang Y and Wang Z (2024) Advancing autonomy through lifelong learning: a survey of autonomous intelligent systems.
Front. Neurorobot. 18:1385778.
doi: 10.3389/fnbot.2024.1385778

The combination of lifelong learning algorithms with autonomous intelligent systems (AIS) is gaining popularity due to its ability to enhance AIS performance, but the existing summaries in related fields are insufficient. Therefore, it is necessary to systematically analyze the research on lifelong learning algorithms with autonomous intelligent systems, aiming to gain a better understanding of the current progress in this field. This paper presents a thorough review and analysis of the relevant work on the integration of lifelong learning algorithms and autonomous intelligent systems. Specifically, we investigate the diverse applications of lifelong learning algorithms in AIS's domains such as autonomous driving, anomaly detection, robots, and emergency management, while assessing their impact on enhancing AIS performance and reliability. The challenging problems encountered in lifelong learning for AIS are summarized based on a profound understanding in literature review. The advanced and innovative development of lifelong learning algorithms for autonomous intelligent systems are discussed for offering valuable insights and guidance to researchers in this rapidly evolving field.

KEYWORDS

artificial intelligence, lifelong learning, algorithm, autonomous intelligent systems, future perspectives

1 Introduction

Autonomous intelligent systems (AIS), including intelligent robots, autonomous vehicles, and similar technologies, have emerged as a frontier direction in the field of artificial intelligence. These systems possess the ability to interact with humans and the environment, enabling them to execute tasks such as perception, planning, decision-making, and control. With the advancement of artificial intelligence, the algorithms employed by AIS for different tasks have transitioned from being model-driven to data-driven approaches. End-to-end AI algorithms based on deep learning, reinforcement learning, and other techniques have gained significant research attention.

However, as the data-driven algorithms rely on the type, scale, and quality of training data, the coherence, generality, and adaptability of the algorithms across different tasks and environments are great challenges. The challenge for AIS concerned in this paper is the ability to remember previous tasks when learning new ones, known as catastrophic forgetting (Shi et al., 2021). Catastrophic forgetting refers to the phenomenon where a neural network loses previously learned information after training on subsequent tasks, resulting in a drastic performance drop on previous tasks (Serra et al., 2018). Therefore, it is crucial to improve the capability of AIS for lifelong learning, which aims to enhance knowledge retention and transfer, thereby addressing the problem of catastrophic forgetting.

Lifelong learning algorithms have made significant progress in dealing with the core problems faced by AIS and mitigating the impact of catastrophic forgetting. Lifelong learning algorithms aim to sequentially acquire proficiency in multiple tasks while pursuing two primary objectives: ensuring that the acquisition of new tasks does not lead to catastrophic forgetting of previously learned knowledge (Zhou and Cao, 2021a), and leveraging prior task knowledge to facilitate the acquisition of novel tasks. Despite numerous achievements in lifelong learning in recent years, there are still evident shortcomings. Firstly, lifelong learning still heavily relies on labeling, which can be costly, troublesome, prone to errors, and impractical for providing persistent human labeling for all future tasks (He et al., 2021). Secondly, adapting to drift in adaptation spaces poses a challenge for lifelong learning. Drift in adaptation spaces arises from uncertainties that impact the quality properties of adaptation options, potentially leading to no adaptation option satisfying the initial set of adaptation goals, thereby damaging system quality (Gheibi and Weyns, 2023). Additionally, the big data problem presents another major challenge. AIS with lifelong learning algorithms must handle the continuous influx of changing data and adapt to learning problems effectively (Yang, 2013).

In this paper, we aim to provide a comprehensive overview of lifelong learning algorithms for autonomous intelligent systems, covering the recent development, related applications, and existing challenges that need to be addressed. Furthermore, we will discuss the future outlook of lifelong learning with autonomous intelligent systems. The main contributions of this paper are as follows:

(1) The thoroughly review and analysis of AIS and lifelong learning, along with the rationale for combining these two fields, are introduced.
(2) Relevant applications of lifelong learning algorithms with AIS are presented to showcase their significant role in different industry applications.
(3) Remaining problems are analyzed, and academic insights into the future trends of AIS Lifelong learning are expounded.

The rest of the paper is organized as follows. Section II elucidates the background information on the emergence and historical milestones of AIS and lifelong learning. Section III presents various applications of lifelong learning algorithms with AIS, highlighting the research status and latest progress. In Section IV, A comprehensive review of issues and challenges in lifelong learning for AIS and the outlook and future trends are discussed. Finally, the main conclusions are given in Section V.

2 The developing lifelong learning and autonomous intelligent systems

2.1 Autonomous intelligent systems

In recent decades, remarkable progress has been made in the development of unmanned systems, ranging from robots to unmanned aerial vehicles (UAVs), unmanned ground vehicles (UGVs), and unmanned marine vehicles (UMVs). What once were programming-based systems have now transformed into automatic unmanned systems and are further advancing toward autonomous intelligent systems (AIS). AIS represents the forefront of artificial intelligence development, characterized by exceptional levels of autonomy and intelligence. By harnessing advanced technologies such as artificial intelligence (AI), big data, and robotics, AIS enables the execution of complex tasks and adaptive decision-making. This section explores the potential applications of AIS across various domains.

2.1.1 Intelligent transportation and autonomous driving

The development of the automobile industry has driven an increased demand for safety and stability in modern transportation. As a result, autonomous driving technology has gained significant traction and is being widely deployed in the market (Xiao, 2022). This technology is revolutionizing intelligent transportation and smart city systems by enhancing the efficiency and safety of transportation networks. It's worth noting that although autonomous driving has recently garnered more attention, the concept of autonomous vehicles dates back several decades, with various activities in this field taking place even further in the past (Khan, 2022).

The first autonomous car was introduced by Tsugawa at the Mechanical Engineering Laboratory in Tsukuba, Japan in the 1970s (OM Group of Companies, 2020). Subsequently, there have been numerous developments and initiatives worldwide. Notably, Ernst Dickmann's vision guided Mercedes Benz in 1980 to achieve speeds of up to 39mph in a controlled environment (Delcker, 2020). With the integration of autonomous driving algorithms, vehicles possess self-navigating capabilities, real-time traffic monitoring, and adaptive route planning based on changing environmental conditions. Furthermore, autonomous driving vehicle enables the efficient management of traffic, congestion control, and the integration of advanced communication and information technologies, thereby facilitating intelligent infrastructure.

However, the utilization of autonomous driving faces significant challenges in complex traffic environments characterized by dynamic and variable scenarios. A key issue lies in perception algorithms encountering the long-tail problem, where rare or unforeseen events pose difficulties for standard algorithms to handle. This challenge becomes even more pronounced in mixed traffic scenarios involving both human-driven and autonomous vehicles. In such settings, algorithms must continually iterate and improve to adapt to the varying and unpredictable nature of the environment (Zhu et al., 2021; Zhou et al., 2022; Li et al., 2023). Therefore, lifelong learning is critical for the development of reliable and safe autonomous systems capable of operating effectively in real-world environments.

2.1.2 Medical healthcare and service robotics

Service robots are typical AIS designed to assist humans, enhancing customer experiences across various industries such as hospitality, logistics, retail, and healthcare (Rajan and Cruz, 2022). With the advancements in AI and IoT technologies, service robots are continuously evolving and becoming more intelligent (Pan et al., 2010). The integration of healthcare and service robotics holds immense promise for improving patient care and enhancing efficiency. Intelligent service robots have the capability to assist in a range of tasks, including patient monitoring, medication dispensing, and patient support, thereby relieving healthcare professionals from repetitive and time-consuming responsibilities. Additionally, intelligent service robots can analyze medical data, provide personalized treatment recommendations, and contribute to remote healthcare services, leading to improved accessibility and quality of care. By leveraging the power of AIS, service robots in healthcare settings can not only streamline processes but also contribute to better patient outcomes. They serve as valuable tools in alleviating the burden on healthcare professionals, enabling them to focus on more complex and critical aspects of patient care. Moreover, AIS-driven analysis of medical data helps generate valuable insights that can inform decision-making and improve treatment strategies (Qu et al., 2021).

However, the integration of intelligent service robots in the field of healthcare also presents certain challenges. One significant challenge is ensuring the safety and reliability of these robots in critical medical environments. As they interact closely with patients, it is essential to address concerns regarding privacy, data security, and potential errors in their operations. Additionally, there is a need for standardized regulations and guidelines to govern the use of service robots in healthcare settings.

Moreover, the complexity and diversity of healthcare scenarios pose challenges for intelligent service robots. Medical environments can be unpredictable, requiring robots to adapt to various situations, handle unexpected events, and effectively communicate with both patients and healthcare professionals. Achieving seamless human-robot interaction and maintaining an appropriate balance between automation and human intervention is crucial in providing high-quality and patient-centric care.

2.1.3 Urban security and UAV

UAV has garnered considerable attention in various military and civilian applications due to their improved stability and endurance (Mohsan et al., 2022). Over the past decade, UAVs have been employed in a wide range of fields, including target detection and tracking, public safety, traffic monitoring, military operations, hazardous area exploration, indoor and outdoor navigation, atmospheric sensing, post-disaster operations, health care, data-sharing, infrastructure management, emergency and crisis management, freight transport, wildfire monitoring and logistics (Hassija et al., 2019). For example, DARPA's "Collaborative Operations in Denied Environment" (CODE) program seeks to enhance the mission capabilities of unmanned aerial vehicles (UAVs) by increasing autonomy and inter-platform collaboration. The United States military has integrated autonomous intelligent unmanned systems into combat through the Project Maven initiative, which employs artificial intelligence algorithms to identify relevant targets in Iraq and Syria. In the domain of urban security, UAV plays a critical role by leveraging AIS's advanced surveillance and analytical capabilities. These intelligent drones enable efficient monitoring of public spaces, early detection of potential threats, and prompt response to emergencies. Moreover, AIS-driven drones enhance search and rescue operations, disaster management, and protection of critical infrastructure while minimizing human risk.

However, several crucial factors hinder the performance of UAVs in urban security. These factors include diverse scenes, stringent man–machine safety requirements, limited availability of training data, and small sample sizes (Carrio et al., 2017; Teixeira et al., 2023). Addressing these challenges is essential to ensure the optimal functioning of UAVs in urban security scenarios. Efforts should be made to develop robust and adaptable AI algorithms that can handle diverse environmental conditions encountered in urban settings. Additionally, ensuring the safety of UAV operations requires stringent regulations and standards for both hardware and software components. Acquiring more extensive and representative training datasets is also necessary to improve the accuracy and reliability of AI models used in UAV systems. Lastly, efforts should be made to address the limitations posed by small sample sizes by leveraging transfer learning techniques and collaborative data sharing initiatives.

2.1.4 Ocean exploration and UMV

AIS contributes significantly to ocean exploration and research through the development of UMV equipped with advanced sensing and navigation capabilities. UMVs integrated with AI algorithms can be used for tasks such as scientific exploration, hydrological surveys, emergency search and rescue, and security patrols (Kingston et al., 2008; Wang et al., 2016). The Monterey Bay Aquarium Research Institute (MBARI) has significantly reduced the human resources required for data analysis by 81% and simultaneously increased the labeling rate tenfold through its Ocean Vision AI program, which trains a vast underwater image database. The autonomous underwater robot, CUREE, developed in collaboration with WHOI, can autonomously track and monitor marine animals, facilitating effective marine management. These wide-ranging applications have contributed to the development of motion control techniques and have produced many interesting results in the literature, such as heading control (Kahveci and Ioannou, 2013), trajectory tracking control (Katayama and Aoki, 2014; Ding et al., 2017), formation control (Li et al., 2018; Liao et al., 2024), and path-following problems (Shen et al., 2019).

The ocean environment presents complex and variable challenges that demand adaptive capabilities from UMV. In the deep-sea environment, UMV encounter various challenges, including changes in underwater terrain, marine biodiversity, and ocean currents. These changes can result in variations in sensor data and diverse appearances of targets. By employing lifelong learning algorithms, unmanned systems can adapt and learn in real-time, enhancing their performance and robustness (Wibisono et al., 2023). Furthermore, deep-sea environments pose limitations in communication bandwidth, latency, and mission execution times. Traditional machine learning algorithms often struggle to adapt to new environments and tasks, as they are typically trained for specific purposes. Lifelong learning algorithms offer a solution by reducing reliance on external resources and human intervention. UMV equipped with these algorithms can autonomously learn and make decisions, increasing their independence and reliability (Wang et al., 2019).

2.1.5 Deep space exploration and spacecraft

Intelligent or autonomous control of an unmanned spacecraft is a promising technology (Soeder et al., 2014). And the ground-based mission control center will no longer be able to help the astronauts diagnose and fix spacecraft issues in real-time due to the longer connection durations associated with deep space exploration, using lifelong learning algorithms, unmanned systems can accumulate experience and knowledge during task execution and reduce reliance on frequent interactions and updates, enhancing their autonomy and adaptability (Jeremy and et.al, 2013). Also, the deep space environment is extremely complex and full of unknown and uncertain factors, such as the landform of the planet's surface, the relationship between celestial bodies, and the atmosphere of the planet. Traditional machine learning algorithms are difficult to pre-train to adapt to all possible situations. Lifelong learning algorithms enable unmanned systems to constantly learn and adapt to new environments and tasks as they explore (Bird et al., 2020). What is more, in deep space exploration missions, unmanned systems typically need to process huge data streams from various sensors and extract useful information from them. Lifelong learning algorithms can help systems automatically discover and learn new features and patterns, thereby improving their perception and understanding (Choudhary et al., 2022). As a result, each vehicle core subsystem will contain inbuilt intelligence to allow autonomous operation for both normal and emergency operations including defect identification and remediation. This extends previous work on creating an autonomous power control (Soeder et al., 2014) which involves the development of control architectures for deep space vehicles (Dever et al., 2014; May et al., 2014) and using software agents (May and Loparo, 2014). As a result, the application of AIS in deep space exploration and spacecraft missions opens up new frontiers for scientific discovery. Intelligent spacecraft equipped with AIS can autonomously navigate, perform complex maneuvers, and adapt to dynamic space environments. Advanced AI-based algorithms enable real-time analysis of vast amounts of space data, autonomous targeting, and intelligent resource allocation, facilitating enhanced mission efficiency and enabling breakthrough discoveries.

In conclusion, the development of unmanned systems has evolved from programming-based to AIS. AIS leverages advanced technologies such as AI, big data, and robotics to enable complex tasks and adaptive decision-making. Across domains including intelligent transportation, healthcare, urban security, ocean exploration, and space missions, AIS demonstrates immense potential for revolutionizing various industries and pushing the boundaries of technological advancements. However, Autonomous intelligent systems require continuous learning to enable their applications in various domains. With the advancements in technologies such as deep learning, reinforcement learning, and large-scale AI models like AIGC (Artificial Intelligence General Cognitive), AISs are moving toward achieving general task learning and lifelong evolution. Establishing a lifelong learning paradigm is crucial for the future development of these autonomous systems. Embracing this paradigm will pave the way for remarkable advancements in the field of autonomous intelligent systems.

Besides the technical perspective, there are actually other angles people should take into consideration to enrich and improve the connotation of autonomous intelligent systems. For one thing, the ethical and social perspective cannot be ignored. Ethically and socially, the deployment of autonomous intelligence systems raises significant questions around accountability, privacy, job displacement, and fairness. The decision-making processes of AIS need to be transparent, explainable, and align with societal values to ensure trust and acceptance. Addressing these concerns involves interdisciplinary research, incorporating insights from ethics, law, and social sciences into the development and governance of AIS. For another thing, autonomous intelligent systems are also closely linked to the Sustainable Development Goals. They have the potential to help address global challenges in environmental protection, health, education and more, such as protecting the environment through intelligent monitoring and management of resources, or improving the quality and accessibility of education through personalized education systems. However, this also requires environmental impact, resource consumption and long-term sustainability to be taken into account when designing and applying autonomous intelligent systems.

2.2 Lifelong learning

Lifelong learning, alternatively known as continuous learning or incremental learning, traces its roots back to the mid-20th century. Early computer scientists and artificial intelligence researchers contemplated ways to enable computer systems to continuously learn and adapt to new knowledge. The adage "one is never too old to learn" holds true and applies equally to AIS.

In 1957, Frank Rosenblatt's perceptron emerged as an early neural network model that introduced the idea of machines improving their ideas and performance gradually through repeated training (Block et al., 1962). The era of artificial intelligence algorithms based on neural networks was begun. But for a long time, neural networks could not handle multiple tasks, nor could they handle dynamic tasks of time series. During the 1990s, the concept and research of transfer learning started to develop, positively influencing the notion of lifelong learning. Transfer learning focused on leveraging previously acquired knowledge for new tasks (Pan et al., 2010). In the 2000s, incremental learning began to emerge in lifelong learning research, enabling AI systems to learn new tasks without sacrificing previously acquired knowledge (Zhou et al., 2022). This approach helps in continuously improving the AI system's performance, adapting to changes in the data distribution, and avoiding catastrophic forgetting. Incremental learning is particularly useful in dynamic environments where new data arrives regularly and the model needs to be continuously updated to maintain its accuracy and relevance. In our dynamically changing world, where new classes appear frequently, fresh users in the authentication system and a machine learning model ought to identify new classes while not forgetting the memory of previous ones (Zhou et al., 2022). If the dataset of old classes is no longer available, directly fine-tuning a deployed model with new classes might bring about the so-called catastrophic forgetting problem in which information about past classes is quickly forgotten (Hinton et al., 2015; Kirkpatrick et al., 2017; Shin et al., 2017). Hence, incremental learning, a framework that enables online learning without forgetting, has been actively investigated (Kang et al., 2022). From the 2000s to 2020s, Researchers have proposed various incremental learning algorithms and techniques to address the challenges associated with learning from evolving data. These algorithms focus on updating the model efficiently (Lv et al., 2019; Tian et al., 2019; Zhao et al., 2021; Ding et al., 2024), handling concept drift (Schwarzerova and Bajger, 2021), managing memory constraints (Smith et al., 2021), and balancing stability and plasticity in the learned knowledge (Wu et al., 2021; Lin et al., 2022; Kim and Han, 2023). Additionally, incremental learning has been explored in different domains, including image classification (Meng et al., 2022; Nguyen et al., 2022; Zhao et al., 2022), natural language processing (Jan Moolman Buys University College University of Oxford, 2017; Kahardipraja et al., 2023), recommender systems (Ouyang et al., 2021; Wang et al., 2021; Ahrabian et al., 2021a), and data stream mining

(Eisa et al., 2022). Researchers have investigated different strategies such as incremental decision trees (Barddal and Fabr'ıcio Enembreck., 2020; Choyon et al., 2020; Han et al., 2023), online clustering (Bansiwala et al., 2021), ensemble methods (Lovinger and Valova, 2020; Zhang J. et al., 2023), and deep learning approaches to tackle incremental learning problems (Ali et al., 2022). Incremental learning enables lifelong learning to constantly learn new data new data while leveraging prior knowledge that continues to be an active research topic (Figure 1).

Lifelong learning plays a crucial role in enhancing the performance of Artificial Intelligence Systems (AIS) due to its powerful capabilities. It enables AIS to continuously update their knowledge and skills, allowing them to effectively handle consecutive tasks in dynamic and evolving environments.

There are three main research methods used in lifelong learning:

- Regularization-based Approach: This method consolidates past knowledge by incorporating additional loss terms that reduce the rate of learning for important weights used in previously learned tasks. By doing so, it minimizes the risk of new task information significantly altering the previously acquired weights (Shaheen et al., 2022). An example of this approach is Elastic Weight Consolidation (EWC), which penalizes weight changes based on task importance, regularizing model parameters and preventing catastrophic forgetting of previous experiences (Febrinanto et al., 2022).
- Rehearsal-based Approach: This method focuses on preserving knowledge by leveraging generative models to replay tasks whenever the model is modified or by storing samples from previously learned tasks in a memory buffer (Faber et al., 2023). One notable approach is Prototype Augmentation and Self-Supervision for Incremental Learning (PASS) (Zhu et al., 2021).
- Model-based Approach: To prevent forgetting, models can be expanded to improve performance, or different models can be assigned to each task. Examples of this approach include Packnet (Mallya and Lazebnik, 2018a) and Dynamically Expandable Representation for Class Incremental Learning (DER) (Yan et al., 2021).

These research methods offer distinct strategies for addressing the challenges associated with lifelong learning in the context of handling consecutive tasks in dynamic and evolving environments. The choice of the most suitable approach depends on specific requirements and circumstances. Ongoing research in the field of lifelong learning continues to explore innovative techniques and approaches to further enhance the performance and adaptability of AIS.

However, the combination of lifelong learning and autonomous intelligent systems poses several challenges due to perceptual cognitive algorithms (Nicolas, 2018; Hadsell et al., 2020), varying tasks (Kirkpatrick et al., 2017; Aljundi et al., 2021), changing environments (Zenke et al., 2017a), and limitations in computing chips (Mallya and Lazebnik, 2018b), control systems (Kober et al., 2013; Andrei et al., 2017), and the diverse range of system types (Kemker and Kanan, 2017; Parisi et al., 2017). Currently, research on this integration is insufficient, and numerous difficulties remain to be addressed. Among these challenges, catastrophic forgetting is a prominent problem wherein previously learned tasks may be forgotten when AIS learns new ones. Consequently, solving this problem holds immense significance and remains a core objective of lifelong learning.

There are three main dimensions to handle catastrophic forgetting:

2.2.1 Knowledge retention

If there is only one model continuously learning different tasks, we naturally expect it not to forget knowledge previously learned when it learns new tasks. In addition, the model is supposed to prevent stopping learning just in order to retain what has been learned at the same time. There are several methods such as Elastic Weight Consolidation (EWC) (Aich, 2021), Synaptic Intelligence (SI) (Zenke et al., 2017b), Memory Aware Synapses (MAS) (Aljundi et al., 2018).

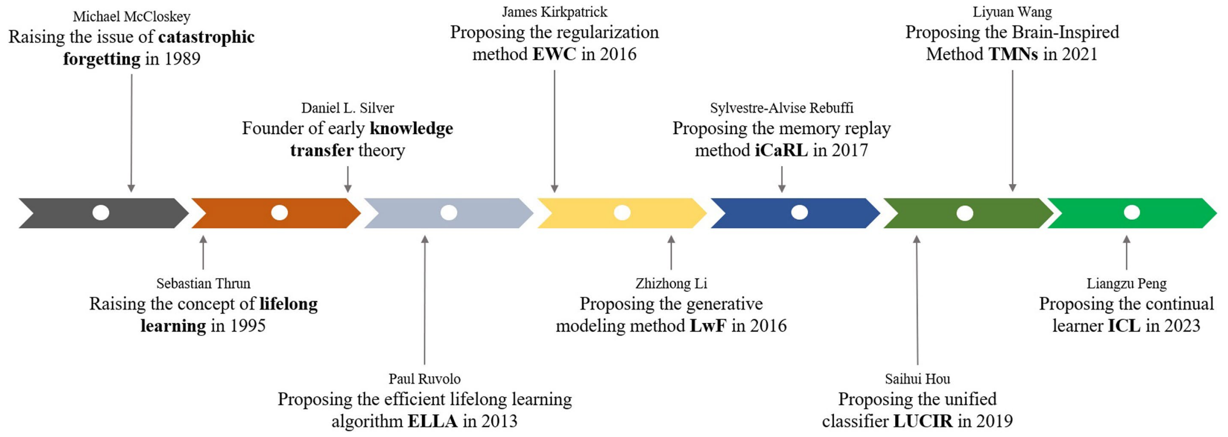

Michael McCloskey
Raising the issue of **catastrophic forgetting** in 1989
Daniel L. Silver
Founder of early **knowledge transfer** theory
James Kirkpatrick
Proposing the regularization method **EWC** in 2016
Sylvestre-Alvise Rebuffi
Proposing the memory replay method **iCaRL** in 2017
Liyuan Wang
Proposing the Brain-Inspired Method **TMNs** in 2021
Sebastian Thrun
Raising the concept of **lifelong learning** in 1995
Paul Ruvolo
Proposing the efficient lifelong learning algorithm **ELLA** in 2013
Zhizhong Li
Proposing the generative modeling method **LwF** in 2016
Saihui Hou
Proposing the unified classifier **LUCIR** in 2019
Liangzu Peng
Proposing the continual learner **ICL** in 2023

FIGURE 1
The history of the development of lifelong learning.

2.2.2 Knowledge transfer

It is expected that models are able to utilize what they have learned to help handle new problems. Related method is Gradient Episodic Memory (GEM) (Lopez-Paz and Ranzato, 2022).

2.2.3 Model expansion

Sometimes, models may be too simple to handle complicated tasks, so it is expected that these models could expand themselves to more complicated ones according to the complexity of problems. Some related methods are Progressive Neural Networks (Rusu et al., 2022), Expert Gate (Aljundi et al., 2017), Net2Net (Chen et al., 2016; Sodhani et al., 2019).

3 Representative applications of lifelong learning for AIS

Nowadays, it is an increasingly popular trend to use lifelong learning algorithms for AIS, which could better improve the performance of these systems. There have been plenty of domains making use of lifelong learning algorithms, here we highlight some representative and contemporary examples below (Figure 2).

3.1 Autonomous driving

The development of autonomous vehicles has advanced quickly in recent years (Han et al., 2023). Modern vehicles are becoming more and more automated and intelligent due to advancements in lifelong learning algorithms, mechanical, and computing technologies (Su et al., 2012). The Institute of Electrical and Electronics Engineers (IEEE) alone produced around 43,000 conference papers and 8,000 journal (including magazine) articles on the subject of autonomous driving in the 5 years between 2016 and 2021 (Chen et al., 2022). Many IT and automotive companies have been attracted to this promising field, such as Baidu Apollo, Google Waymo. And by 2021, Waymo's autonomous vehicles have driven more than 20 million miles on the road, demonstrating the reliability and safety of the technology of autonomous driving. As a result, in the near future, different types of AVs are expected to be fully commercialized, with a significant impact on all aspects of our lives (Su et al., 2012).

The most challenging problem autonomous driving currently faces is to adapt to novel driving scenarios, especially in complex and mixed traffic environments, and react properly and rapidly in time. As a result, autonomous driving is particularly in need of the combination of lifelong learning algorithms. So in the section below, different frames of lifelong learning in some crucial fields of autonomous driving are explained.

3.1.1 Lane changing

Lane changing is one of the largest challenges in the high-level decision-making of autonomous vehicles (AVs), especially in mixed and dynamic traffic scenarios, where lane changing has a significant impact on traffic safety and efficiency. In recent years, the application of lifelong learning to lane-changing decision-making in AVs has been widely explored with encouraging results. However, most of these studies have focused on single-vehicle environments, and lane-changing in situations where multiple AVs coexist with human-driven vehicles has received little attention (Zhou et al., 2022), which should be paid more attention. In this regard, Ref. (Zhou et al., 2022) proposes a multi-agent advantage actor-critic method which uses a novel local reward design and parameter sharing scheme to formulate the lane changing decision of multiple AVs in a mixed traffic highway environment as a multi-agent lifelong learning problem using a lifetime learning algorithm.

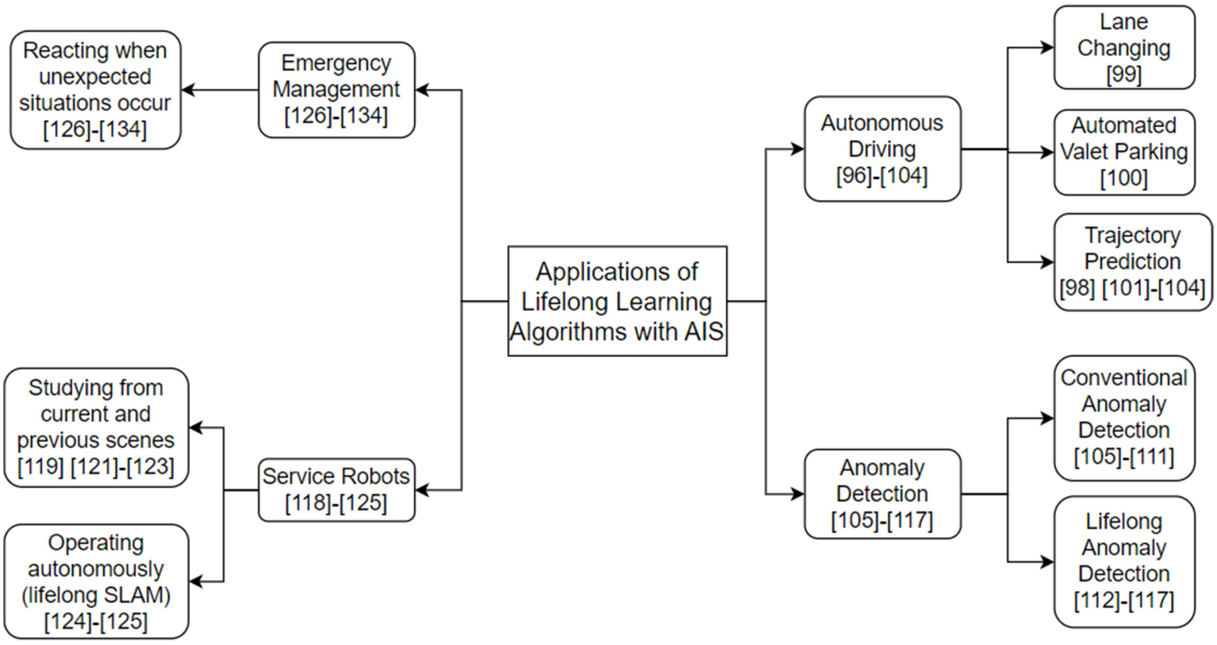

Reacting when unexpected situations occur [126]-[134]
Emergency Management [126]-[134]
Applications of Lifelong Learning Algorithms with AIS
Autonomous Driving [96]-[104]
Lane Changing [99]
Automated Valet Parking [100]
Trajectory Prediction [98] [101]-[104]
Anomaly Detection [105]-[117]
Conventional Anomaly Detection [105]-[111]
Lifelong Anomaly Detection [112]-[117]
Service Robots [118]-[125]
Studying from current and previous scenes [119] [121]-[123]
Operating autonomously (lifelong SLAM) [124]-[125]

FIGURE 2
The taxonomy of technologies related to applications of lifelong learning algorithms for AIS.

3.1.2 Automated valet parking

Automated valet parking (AVP) allows human drivers to park their cars in a drop-off zone (e.g., a parking garage entrance). These cars can independently perform autonomous driving tasks from the parking area to a designated parking space. AVP can greatly improve driver convenience, and is seen as an entry point for the promotion of AVs. And high-precision indoor positioning service is unavoidable in AVP. However, existing wireless indoor positioning technologies, including Wi-Fi, Bluetooth, and ultra-wideband (UWB), have a tendency to degrade significantly with the increase of working time and the change of building environments (Zhao et al., 2023). To handle this problem, a data-driven and map-assisted indoor positioning correction model has been proposed to improve the positioning accuracy for the infrastructure-enabled AVP system recently by a research team from Tongji University, Shanghai, China (for details refer to Ref. (Zhao et al., 2023)). In order to sustain the lifelong performance, the model is updated in an adversarial manner using crowdsourced data from the on-board sensors of fully instrumented autonomous vehicles (Zhao et al., 2023).

3.1.3 Trajectory prediction

Accurate trajectory prediction of vehicles is the key to reliable autonomous driving. Adapting to changing traffic environments and implementing lifelong trajectory prediction models are crucial in order to maintain consistent vehicle performance across different cities. In real applications, intelligent vehicles equipped with autonomous driving systems should travel on different roadways, cities and even countries. The system needs to properly forecast the future trajectories of the surrounding vehicles and adapt to the diverse distribution of their motion and interaction pattern in order to safely guide the vehicle. In order to achieve this, the system must constantly acquire new information about developing traffic conditions while retaining its previous understanding. Furthermore, the system cannot afford to store a significant amount of trajectory data due to its restricted storage resources (Bao et al., 2021). So, in order to perform well on all processed tasks, it is necessary to keep lifelong learning with restricted storage resource. As a consequence, in a bid to achieve lifelong trajectory prediction, a new framework based on conditional generative replay is proposed by the research team from the University of Science and Technology of China (USTC), which handles the problem of catastrophic forgetting due to different types of traffic environments and improve the precision and efficiency of vehicle trajectory prediction (Bao et al., 2021).

At the moment, autonomous vehicles are not perfect in their operation (Chen et al., 2022), as evidenced by some accidents caused by autonomous driving vehicles in recent years, in which safety drivers were unable to prevent the accidents from occurring, resulting in the loss of multiple lives, thus bringing about these mournful aftermaths which could have been prevented. Obviously, in terms of performance, autonomous vehicle systems are still far from the visual systems of humans or animals (Chen et al., 2020). It is necessary to find novel solutions, such as bio-inspired visual sensing, multi-agent collaborative perception, and control capabilities that emulate biological systems' operational principles (Tang et al., 2021). It is predicted that after reaching increasing degrees of robotic autonomy and vehicle intelligence, autonomous driving will become sufficiently safe and dependable by 2030 to replace the majority of human driving (Litman, 2021).

3.2 Anomaly detection

Anomaly detection is the task of finding anomalous data instances, which therefore represents deviations from the normal conditions of a process (Aggarwal, 2017). In many fields and real-world applications, such as network traffic invasions (Faber et al., 2021), aberrant behavior in cyber-physical systems like smart grids (Corizzo et al., 2021), or flaws in manufacturing processes (Alfeo et al., 2020), the ability to identify abnormal behavior is crucial.

Examples of relevant techniques for detecting anomalies in one-class learning are: (i) Autoencoder, a model based on neural network reconstruction; (ii) One-Class Support Vector Machine, which provides anomaly scores by contrasting new data with the decision boundary based on hyperplanes.; (iii) Local Outlier Factor, which provides an anomaly score that is derived from the ratio of the new data samples' local density to the average local density of its closest neighbors; (iv) Isolation Forest, which offers tree ensembles and calculates the new samples' anomaly score by measuring the distance from the root to the leaf; (v) Copula-based anomaly detection, which draws conclusions about the level of "extremeness" of data samples by using tail probabilities (Goldstein and Uchida, 2016; Li et al., 2020; Lesouple et al., 2021).

However, although these methods have been established and perform well in many scenarios, due to the catastrophic forgetting, the performance of the anomaly detection system is affected negatively when previous circumstances reoccur. For this reason, lifelong anomaly detection is supposed to be applied to balance between knowledge transferring and knowledge retention. Since many real-world domains are characterized by both recurrent conditions and dynamic, rapidly evolving situations, lifelong anomaly detection may out to be quite advantageous in these kinds of environments. This feature necessitates model characteristics that promote concurrent learning and adaptability (Faber et al., 2023). And several recent research efforts have begun to address the problem of lifelong anomaly detection. Examples include using meta-learning to estimate parameters for numerous tasks in one-class image classification (Frikha et al., 2021), transfer learning for anomaly detection in videos (Doshi and Yilmaz, 2020), and change-point detection in conjunction with memory arrangement (Corizzo et al., 2022). Particularly, in the field of autonomous driving, an effective collaborative anomaly detection methodology known as ADS-Lead was proposed to safeguard the lane-following mechanism of ADSs. It has a unique transformer-based one-class classification algorithm to detect adversarial image examples (traffic sign and lane identification threats) as well as time series anomalies (GPS spoofing threat) (Han et al., 2023). In addition, to enhance the anomaly detection performance of models, an active lifetime anomaly detection framework was provided for class-incremental scenarios that supports any memory-based experience replay mechanism, any query strategy, and any anomaly detection model (Faber et al., 2022).

Figure 3 illustrates a typical scenario comparing conventional anomaly detection with model updating with lifelong anomaly detection. In contrast to conventional anomaly detection, which continuously updates the model and causes detection delays, or false predictions, until the new task is incorporated into the model, lifetime anomaly detection in the second iteration does not require model updates following a recurrence of each work. Furthermore, in a 100-iteration scenario, only 4 model updates would be needed for

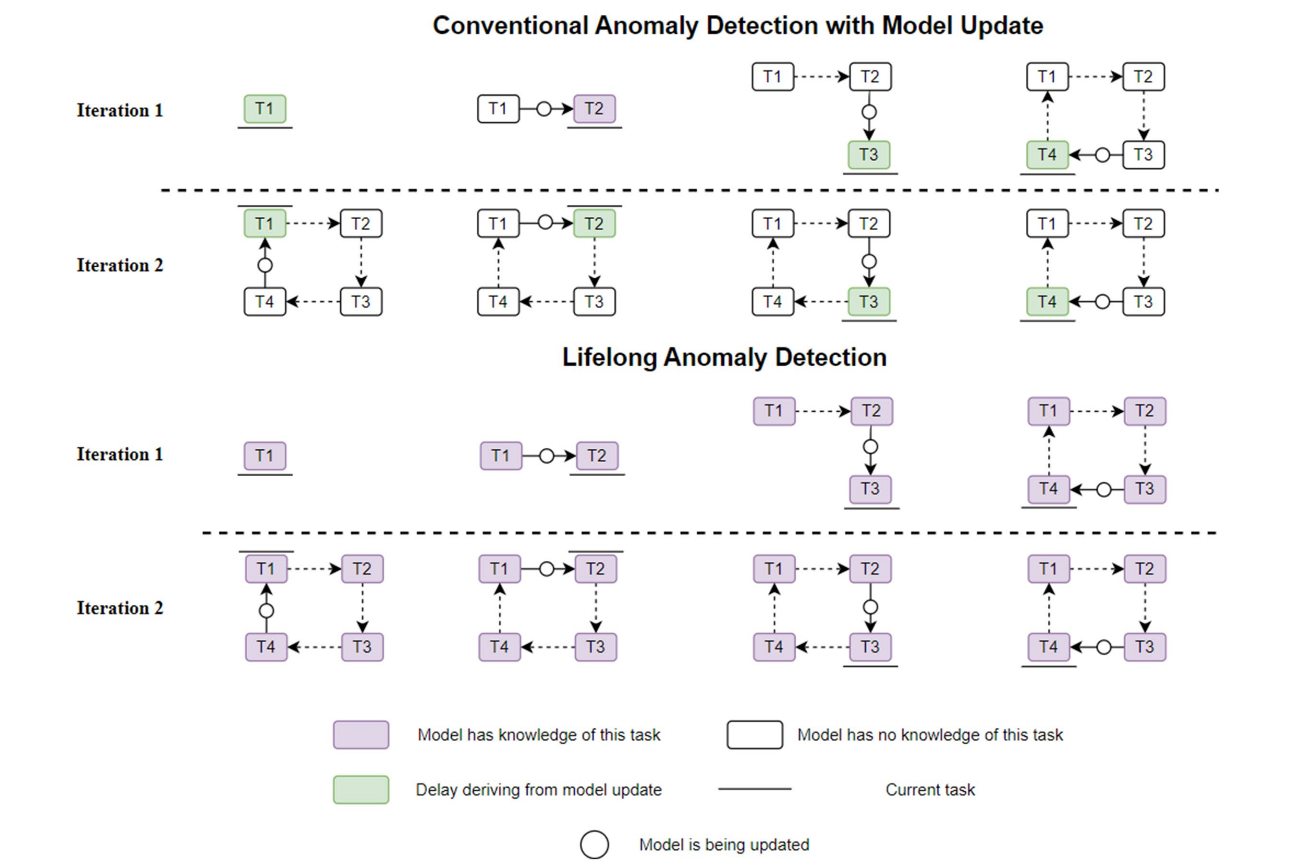

Conventional Anomaly Detection with Model Update
Iteration 1
Iteration 2
T1
T2
T3
T4
Lifelong Anomaly Detection
Iteration 1
Iteration 2
Model has knowledge of this task
Model has no knowledge of this task
Delay deriving from model update
Current task
Model is being updated

FIGURE 3
A scenario with four recurring tasks (Comparisons between conventional and lifelong anomaly detection).

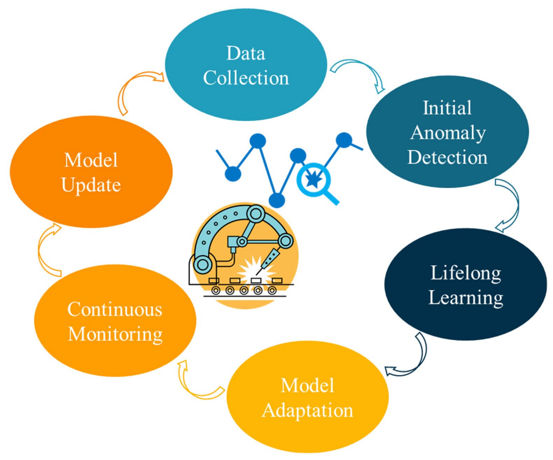


FIGURE 4
The core process of lifelong anomaly detection.

lifetime anomaly detection, as opposed to 400 model updates for traditional anomaly detection, which results in detection delays. It could be used to map a wide range of recurring real-world scenarios, such as human activity sequences, geophysical phenomena like weather patterns, and cyber-physical system operating conditions (Faber et al., 2023; Figure 3).

The core process of lifelong anomaly detection involves several key steps, as depicted in Figure 4. These steps include data collection, initial anomaly detection, lifelong learning, model adaptation, continuous monitoring, model update, and the repetition of the process. The first step is data collection, wherein data is gathered from multiple sources, such as network traffic, smart grids, and manufacturing processes. Following data collection, initial anomaly detection techniques, such as Autoencoders, Support Vector Machines, Local Outlier Factor, and Isolation Forests, are employed to conduct preliminary anomaly detection. Subsequently, lifelong learning takes place, whereby new data is integrated into the model while existing knowledge is updated and retained. Model adaptation is then performed based on the new data, which may involve applying techniques like meta-learning, transfer learning, or change point detection with memory organization. Continuous monitoring of the data for anomalies is carried out to ensure timely detection. To maintain the model's effectiveness, periodic model updates are performed by refreshing it with new data and employing advanced techniques. This entire process is repeated cyclically, encompassing both data collection and model updating stages.

3.3 Service robots

Depending on the continuous learning mechanism for a variety of various robotic tasks, lifelong machine learning has drawn intriguing academic interests in the field of robotics (Dong et al., 2022). And past research has identified lifelong learning as a critical capability for service robots. Creating an artificial "lifelong learning" agent that can construct a cultivated understanding of the world from the current scene and their prior knowledge through an autonomous lifelong system is one of the big ambitions of robotics (She et al., 2020). According to a report by Allied Market Research, the global service robot market is valued at $21.084 billion in 2020 and is expected to reach $293.087 billion by 2032, with a CAGR of 24.3% from 2023 to 2032. Moreover, the number of new startups named after service robots accounts for 29% of all U.S. robotics companies. Those data, among other similar figures, remark the development in the service robots area (Gonzalez-Aguirre et al., 2021). Service robots are mostly tasked with helping humans in the home environment, and they must handle a wide variety of objects. These objects are dependent on the particular environment (e.g., bedroom, toilet, balcony), the human being supported (e.g., kids, elderly people, disabled people). It is practically impossible to prepare all possible objects at the time of or prior to the deployment of the robot. Therefore, the robots will need to adjust to new objects and different ways of perceiving things throughout their lives (Niemueller, 2013). Despite these challenges, we want these robots to notice us and show adaptive behavior when they are on a mission. When a robot is given negative feedback when vacuuming while someone is watching TV, it should be able to recognize this as a new context and adjust its behavior accordingly in similar spatial or social contexts. For example, when people are reading books, the robot should be able to connect this scenario to the one it has previously encountered and cease vacuuming (Irfan et al., 2021). Another example is when service robots engage in language teaching, they may encounter variations in language environments and user learning needs. In such cases, it is imperative for service robots to achieve self-learning and improvement by monitoring user feedback, autonomously exploring language environments, and utilizing natural language processing techniques. Only through these means can they better provide personalized language learning support and practical opportunities for users, thus enhancing teaching proficiency and efficiency (Kanero et al., 2022).

Another aspect of lifelong learning applied to robots is the ability to function independently for lengthy periods of time in dynamic, constantly-changing surroundings. For example, in a domestic scene, where most objects are likely to be movable and interchangeable, the visual character of the same place may differ markedly over successive days. To deal with this situation, a term lifelong SLAM has been in use to address SLAM problems in environments that have been changing over time, improving the robustness and accuracy of pose estimation of robots (Shi et al., 2020). Lifelong SLAM takes into account a robot's long-term operations, which involve repeatedly visiting previously mapped places in dynamic surroundings. In lifetime SLAM, we make the assumption that a region is constantly mapped over an extended period of time, rather than only once (Kurz et al., 2021). Compared to classical SLAM methods, however, there exist a lot of challenges (Shi et al., 2020):

- Changed viewpoints - the robot may look at the same scene or items from several angles.
- Changed things - the objects may have been changed when reentering a place that was previously observed by the robot.
- Changed illumination - there could be a significant change in illumination.
- Dynamic objects - There could be objects in the scene that are moving or changing.
- Degraded sensors - unpredictable sensor noises and calibration errors could result from a variety of factors, including mechanical strain, temperature changes, dirty or damp lenses, etc.

To address these challenges, the operational flow of the lifetime service robot is shown in Figure 5.

3.4 Emergency management

In recent years, machine learning algorithms have made great strides in enabling autonomous agents to learn through observation and sensor feedback how to carry out tasks in complex online environments. In particular, recent developments in deep neural network-based lifelong learning have demonstrated encouraging outcomes in the creation of autonomous agents that can interact with their surroundings in a variety of application domains (Arulkumaran et al., 2017), including learning to play games (Brown and Sandholm, 2017; Xiang et al., 2021), generating optimal control policies for robots (Jin et al., 2017; Pan et al., 2017), natural language processing and speech recognition (Bengio et al., 2015), body emotion understanding (Sun and Wu, 2023), as well as choosing the best trades in light of the shifting market conditions (Deng et al., 2017). The agent gradually learns the best course of action for the assigned task by seeing how its actions result in rewards from these encounters.

These methods are effective when it can be presumed that every event that occurs during deployment is a result of the same distribution that the agent was trained on. However, agents that must operate for extended periods of time in complex, real-world environments may be subject to unforeseen circumstances beyond the distribution for which they were designed or trained, due to changes in the environment. For instance, a construction site worker may unintentionally place a foreign object—like their hand–inside the workspace of a vision-guided robot arm, which must then react to prevent harm or damage. Similarly, an autonomous driving car may come across significantly distorted lane markings that it has never encountered before and must decide how to continue driving safely. In such unexpected and novel situations, the agent's strategy will not apply, leading to the possibility of the agent taking unsafe actions. And that is what makes emergency management crucial.

The purpose of emergency management is to provide autonomous agents with the ability to respond to unforeseen situations that are different from what they are trained or designed to handle. Therefore, a lifelong data-driven response-generation system must be developed to tackle this problem. It enables an agent to handle new scenarios without depending on the reliability of pre-existing models, safe states, and recovery strategies created offline or from prior experiences, or on their accuracy. The main finding is that, when needed, uncertainty in environmental observations may be used to inform the creation of quick, online reactions that effectively avoid threats and allow the agent to carry on operating and learning in its surroundings (Maguire et al., 2022). As is shown in Figure 6, the core process of emergency management has a close relationship with lifelong learning algorithms, it keeps learning and adapting.

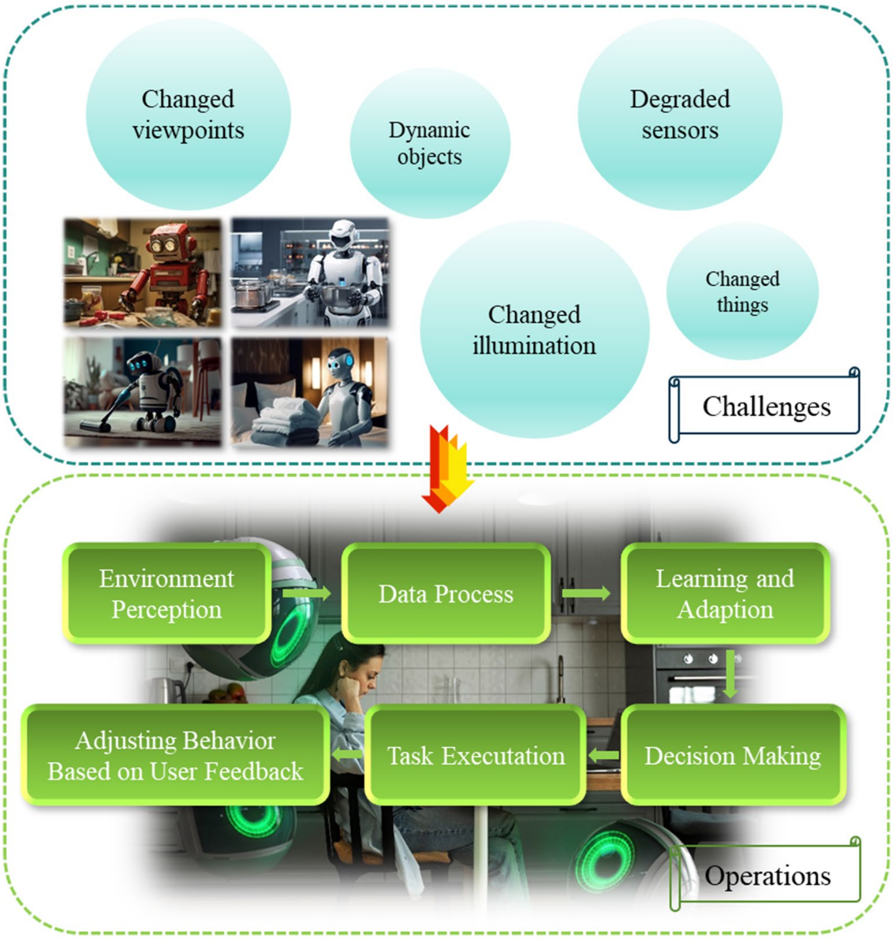


FIGURE 5
The operational flow of the lifetime service robot.

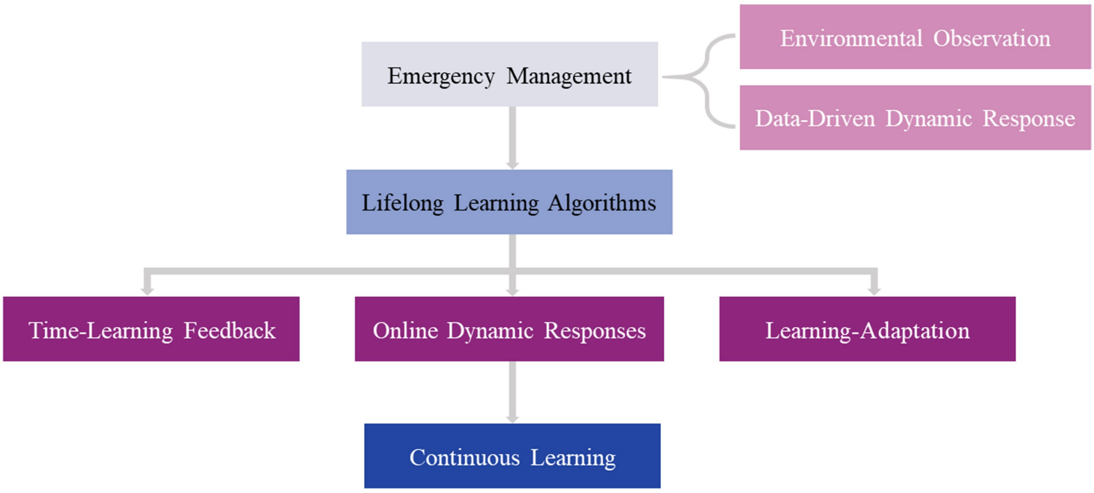


FIGURE 6
The emergency management process based on lifelong learning.

4 Outlook

Lifelong learning with AIS has made significant progress in recent decades. And graph lifelong learning is emerging as an important area in AI research and applications. Graph lifelong learning involves applying lifelong learning principles to graph-based data structures and algorithms. This approach aims to enable systems to continuously learn and adapt from a stream of graph data over time. There are many kinds of graph lifelong learning algorithms, and there exist several differences between these methods, which suit different situations. Each method may have its own approach and principle to cope with problems, as can be seen in Table 1.

The key challenge in graph lifelong learning is to efficiently update and refine the model as new data arrives, without forgetting previously learned information (Galke et al., 2023). In addition, dynamic nature of graphs also brings problems for the reasons that graph data is often dynamic, such as social networks or knowledge graphs. Models need to adapt to these changes while maintaining the validity of past learning (Galke et al., 2023). Graph lifelong learning is a rapidly growing field that proposes new solutions for how intelligent systems can continuously learn and adapt to changing environments. With further research, this field is expected to solve existing challenges and provide strong support for the continued development and application of intelligent systems.

Besides the development of graph lifelong learning, several trends and directions can be observed in the relationship between lifelong learning algorithms and AIS. Firstly, multi-modal learning will play a crucial role as autonomous systems learn from diverse sensors and data sources, including visual, auditory, textual, and sensor data. This integration will greatly enhance the system's perception and understanding capabilities. Secondly, an important aspect is self-improvement learning, where the system autonomously assesses its performance, identifies weaknesses, and automatically adjusts and improves its algorithms and models to enhance efficiency and accuracy. Furthermore, cross-domain transfer of knowledge and experience becomes a possibility. The system will be able to transfer learned knowledge from one domain to another, thereby enhancing its problem-solving abilities across different domains. What is more, lifelong learning with AIS can also be developed and applied in the area of education, especially in English teaching and learning. According to Grand View Research, the AI market in education is expected to reach $13.3 billion by 2025. Its diversity is able to change the form of language education to a certain extent, making it continuously transform from the original, traditional, and monotonous form to a dimensional, dynamic, and multi-spatial form, providing a personalized learning experience based on individual needs and preferences (Hwang et al., 2020). Although there has been little research on how lifelong learning can enhance English teaching and learning through AIS so far, it can benefit this area without doubt (Gao, 2021; Pikhart, 2021; Klimova et al., 2022).

Concerning lifelong learning algorithms themselves, incremental learning should receive more attention. Improving the efficiency and stability of incremental learning becomes crucial, enabling the system to retain previous knowledge while learning new tasks. Additionally, self-supervised learning methods will gain prominence. These techniques allow systems to learn from unlabeled data, reducing reliance on extensive labeled data and opening up opportunities for continuous learning. Overall, these trends and directions highlight the importance of multi-modal learning, self-improvement learning, cross-domain transfer, efficient incremental learning, and self-supervised learning in advancing the field of lifelong learning algorithms for AIS.

5 Conclusion

In this paper, we have extensively discussed the relationship between lifelong learning algorithms and autonomous intelligent systems. We have demonstrated the specific applications of lifelong learning algorithms in various domains such as autonomous driving, anomaly detection, service robotics, and emergency management. It is found that current research has made certain progress in addressing

TABLE 1 Graph lifelong learning method comparison.

Methods	Approach			
	Architectural	Rehearsal	Regularization	Reference
Feature Graph Networks	Yes	No	No	Sarlin et al. (2020) and Zhou et al. (2022)
Hierarchical Prototype Networks	Yes	No	No	Li et al. (2023) and Zhang et al. (2023a)
Experience Replay GNN Frame work	No	Yes	No	Ahrabian et al. (2021a) and Zhou and Cao (2021b)
Lifelong Open-world Node Classification	No	Yes	No	Galke et al. (2021) and Zhang et al. (2022)
Disentangle-based Continual Graph Representation Learning	No	No	Yes	Kou et al. (2020) and Zhang et al. (2023b)
Graph Pseudo Incremental Learning	No	No	Yes	Tan et al. (2022) and Su et al. (2023)
Topology-aware Weight Preserving	No	No	Yes	Natali et al. (2020) and Liu et al. (2021)
Translation-based Knowledge Graph Embedding	No	No	Yes	Yoon et al. (2016) and Li et al. (2023)
Continual GNN	No	Yes	Yes	Han et al. (2020) and Wang et al. (2020)
Lifelong Dynamic Attributed Network Embedding	Yes	Yes	Yes	Li et al., 2017, Yoon et al. (2017), and Liu et al. (2021)

the catastrophic forgetting problem of complex scenarios and multitasking under long time sequences. However, challenges such as activation drift, inter-task confusion, and excessive neural resources still persist. In light of this, we particularly emphasize the significance and potential of advancing lifelong learning through graphical approaches, while pointing out that multimodal learning and methods like cross-domain transfer are pivotal references for future advancements in AIS lifelong learning algorithms. Among these, the integration of robot vision and tactile perception is recognized as a key challenge to enhance robot performance and efficiency. To conclude, lifelong learning proves to be a reliable and efficient method for advancing autonomous intelligent systems. Future research efforts should focus on developing fully autonomous and secure learning frameworks that offer superior performance while reducing the need for excessive supervision, training time, and resources.

Author contributions

DZ: Conceptualization, Methodology, Resources, Writing – review & editing. QB: Investigation, Visualization, Writing – original draft. ZZ: Conceptualization, Funding acquisition, Methodology, Resources, Writing – review & editing. YZ: Investigation, Visualization, Writing – original draft. ZW: Funding acquisition, Resources, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported in part by the National Key Research and Development Program of China (t 2021YFE0193900 and 2020AAA0108901), in part by the Fundamental Research Funds for the Central Universities (22120220642).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aggarwal, C. C. (2017). *An introduction to outlier analysis*, Berlin: Springer, pp. 1–34.

Ahrabian, K., Xu, Y., Zhang, Y., Wu, J., Wang, Y., and Coates, M. (2021a). *Structure aware experience replay for incremental learning in graph-based recommender systems*. CIKM '21: The 30th ACM International Conference on Information and Knowledge Management.

Aich, A. (2021). Elastic weight consolidation(EWC): nuts and bolts. *arXiv* 2021:004093v1. doi: 10.48550/arXiv.2105.04093

Alfeo, A. L., Cimino, M. G., Manco, G., Ritacco, E., and Vaglini, G. (2020). Using an autoencoder in the design of an anomaly detector for smart manufacturing. *Pattern Recogn. Lett.* 136:8. doi: 10.1016/j.patrec.2020.06.008

Ali, R., Hardie, R. C., Narayanan, B. N., and Kebede, T. M. (2022). IMNets: deep learning using an incremental modular network synthesis approach for medical imaging applications. *Appl. Sci.* 12:5500. doi: 10.3390/app12115500

Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., and Tuytelaars, T. (2018). Memory aware synapses: learning what (not) to forget. *arXiv* 2018:09601v4. doi: 10.48550/arXiv.1711.09601

Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., and Tuytelaars, T. (2021). *Memory aware synapses: learning what (not) to forget*. Proceedings of the European Conference on Computer Vision (ECCV), pp. 144–161.

Aljundi, R., Chakravarty, P., and Tuytelaars, T. (2017). Expert gate: lifelong learning with a network of experts. *arXiv* 2017:06194v2. doi: 10.48550/arXiv.1611.06194

Andrei, A., Rusu, M. V., Rothörl, T., Heess, N., Pascanu, R., and Hadsell, R. (2017). *Sim-to-real robot learning from pixels with progressive nets*. Proceedings of the 1st Annual Conference on Robot Learning, PMLR, No. 78, pp. 262–270.

Arulkumaran, K., Deisenroth, M. P., Brundage, M., and Bharath, A. A. (2017). Deep reinforcement learning: a brief survey. *IEEE Signal Process. Mag.* 34, 26–38. doi: 10.1109/MSP.2017.2743240

Bansiwala, R., Gosavi, P., and Gaikwad, R. (2021). Continual learning for food recognition using class incremental extreme and online clustering method: self-organizing incremental neural network. *Int. J. Innov. Eng. Sci.* 6, 36–40. doi: 10.46335/IJIES.2021.6.10.7

Bao, P., Chen, Z., Wang, J., Dai, D., and Zhao, H. (2021). Lifelong vehicle trajectory prediction framework based on generative replay. *arXiv* 2021:0751. doi: 10.1109/TITS.2023.3300545

Barddal, J. P., and Enembreck, F. (2020). Regularized and incremental decision trees for data streams. *Ann. Telecommun.* 75, 493–503. doi: 10.1007/s12243-020-00782-3

Bengio, S., Vinyals, O., Jaitly, N., and Shazeer, N. (2015). *Scheduled sampling for sequence prediction with recurrent neural networks*. In: Proceedings of the 28th international conference on neural information processing systems. MIT Press, Cambridge, MA, pp. 1171–1179.

Bird, J., Colburn, K., Petzold, L., and Lubin, P. (2020). Model optimization for deep space exploration via simulators and deep learning. *ArXiv* 2020:14092. doi: 10.48550/arXiv.2012.14092

Block, H. D., Knight, B. W., and Rosenblatt, F. (1962). Analysis of a four-layer series-coupled perception. II*. *Rev. Mod. Phys.* 34, 135–142. doi: 10.1103/RevModPhys.34.135

Brown, N., and Sandholm, T. (2017). *Libratus: the superhuman AI for no-limit poker. In Proceedings of the twenty-sixth international joint conference on artificial intelligence (IJCAI-17)*. IJCAI Organization, Menlo Park, Calif, pp. 5226–5228.

Carrio, A., Sampedro, C., Rodriguez-Ramos, A., and Campoy, P. (2017). A review of deep learning methods and applications for unmanned aerial vehicles. *J Sens* 2017:3296874. doi: 10.1155/2017/3296874

Chen, G., Cao, H., Conradt, J., Tang, H., Rohrbein, F., and Knoll, A. (2020). Event-based neuromorphic vision for autonomous driving: a paradigm shift for bioinspired visual sensing and perception. *IEEE Signal Process. Mag.* 37, 34–49. doi: 10.1109/MSP.2020.2985815

Chen, T., Goodfellow, I., and Shlens, J. (2016). Net2Net: accelerating learning via knowledge transfer. *arXiv* 2016:05641v4. doi: 10.48550/arXiv.1511.05641

Chen, J., Sun, J., and Wang, G. (2022). From unmanned systems to autonomous intelligent systems. *Engineering* 12, 16–19. doi: 10.1016/j.eng.2021.10.007

Choudhary, K., DeCost, B., Chen, C., Jain, A., Tavazza, F., Cohn, R., et al. (2022). Recent advances and applications of deep learning methods in materials science. *NPJ Comput. Mater.* 8:59. doi: 10.1038/s41524-022-00734-6

Choyon, A., Md, A. R., Hasanuzzaman, D. M. F., and Shatabda, S. (2020). Incremental decision trees for prediction of adenosine to inosine RNA editing sites. *F1000Research* 9:11. doi: 10.12688/f1000research.22823.1

Corizzo, R., Baron, M., and Japkowicz, N. (2022). Cpdga: change point driven growing auto-encoder for lifelong anomaly detection. *Knowl. Based Syst.* 2022:108756. doi: 10.1016/j.knosys.2022.108756

Corizzo, R., Ceci, M., Pio, G., Mignone, P., and Japkowicz, N. (2021). *Spatially-aware autoencoders for detecting contextual anomalies in geo-distributed data*. In: International Conference on Discovery Science, Springer, pp. 461–471.

Delcker, J. (2020). *The man who invented the self-driving Car (in 1986)*. Available at: www.politico.eu/article/delf-driving-car-born-1986-ernst-dickmanns-merc%edes/.

Deng, Y., Bao, F., Kong, Y., Ren, Z., and Dai, Q. (2017). Deep direct reinforcement learning for financial signal representation and trading. *IEEE Trans. Neural Netw. Learn. Syst.* 28, 653–664. doi: 10.1109/TNNLS.2016.2522401

Dever, T. P., Trase, L. M., and Soeder, J. F. (2014). *Application of autonomous spacecraft power control technology to terrestrial microgrids*. AIAA-2014-3836, AIAA propulsion and energy forum, 12th international energy conversion engineering conference, Cleveland, OH.

Ding, S., Feng, F., He, X., Liao, Y., Shi, J., and Zhang, Y. (2024). Causal incremental graph convolution for recommender system retraining. *IEEE Trans. Neural Netw. Learn. Syst.*, 1–11. doi: 10.1109/TNNLS.2022.3156066

Ding, L., Xiao, L., Liao, B., Lu, R., and Peng, H. (2017). An improved recurrent neural network for complex-valued Systems of Linear Equation and its Application to robotic motion tracking. *Front. Neurorobot.* 11:45. doi: 10.3389/fnbot.2017.00045

Dong, J., Cong, Y., Sun, G., and Zhang, T. (2022). Lifelong robotic visual-tactile perception learning. *Pattern Recogn.* 121:108176. doi: 10.1016/j.patcog.2021.108176

Doshi, K., and Yilmaz, Y. (2020). *Continual learning for anomaly detection in surveillance videos*. In: Proceedings of the IEEE/CVF CVPR workshops, pp. 254–255.

Eisa, A., EL-Rashidy, N., Alshehri, M. D., El-Bakry, H. M., and Abdelrazek, S. (2022). Incremental learning framework for mining big data stream. *Comput. Mater. Contin.* 2022:342. doi: 10.32604/cmc.2022.021342

Faber, K., Corizzo, R., Sniezynski, B., and Japkowicz, N.. (2022). *Active lifelong anomaly detection with experience replay*. 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA).

Faber, K., Corizzo, R., Sniezynski, B., and Japkowicz, N. (2023). Lifelong learning for anomaly detection: new challenges, perspectives, and insights. *arXiv* 2023:07557v1. doi: 10.48550/arXiv.2303.07557

Faber, K., Faber, L., and Sniezynski, B. (2021). *Autoencoder-based ids for cloud and mobile devices*. In: 2021 IEEE/ACM 21st CCGrid, IEEE, pp. 728–736.

Febrinanto, F. G., Xia, F., Moore, K., Thapa, C., and Aggarwal, C. (2022). Graph lifelong learning: a survey. *arXiv* 2022:10688v2. doi: 10.48550/arXiv.2202.10688

Frikha, A., Krompaß, D., and Tresp, V. (2021). *Arcade: a rapid continual anomaly detector*. In: 2020 25th international conference on pattern recognition (ICPR), IEEE, pp. 10449–10456.

Galke, L., Franke, B., Zielke, T., and Scherp, A.. (2021). *Lifelong learning of graph neural networks for open-world node classification*. 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, pp. 1–8.

Galke, L., Vagliano, I., Franke, B., Zielke, T., Hoffmann, M., and Scherp, A. (2023). Lifelong learning on evolving graphs under the constraints of imbalanced classes and new classes. *Neural Netw.* 164, 156–176. doi: 10.1016/j.neunet.2023.04.022

Gao, J. (2021). Exploring the feedback quality of an automated writing evaluation system pigai. *Int. J. Emerg. Technol. Learn.* 16, 322–330. doi: 10.3991/ijet.v16i11.19657

Gheibi, O., and Weyns, D. (2023). Dealing with drift of adaptation spaces in learning-based self-adaptive systems using lifelong self-adaptation. *arXiv* 2023:02658. doi: 10.48550/arXiv.2211.02658

Goldstein, M., and Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLoS One* 11:e0152173. doi: 10.1371/journal.pone.0152173

Gonzalez-Aguirre, J. A., Osorio-Oliveros, R., Rodríguez-Hernández, K. L., Lizárraga-Iturralde, J., Morales Menendez, R., Ramírez-Mendoza, R. A., et al. (2021). Service robots: trends and technology. *Appl. Sci.* 11:10702. doi: 10.3390/app112210702

Hadsell, R., Rao, D., Rusu, A. A., and Pascanu, R. (2020). Embracing change: continual learning in deep neural networks. *Trends Cogn. Sci.* 24, 1028–1040. doi: 10.1016/j.tics.2020.09.004

Han, Z., Ge, C., Member, I. E. E. E., Bingzhe, W., Liu, Z., and Senior Member, I. E. E. E. (2023). Lightweight privacy-preserving federated incremental decision trees. *IEEE Trans. Serv. Comput.* 16:1. doi: 10.1109/TSC.2022.3195179

Han, Y., Karunasekera, S., and Leckie, C. (2020). Graph neural networks with continual learning for fake news detection from social media. *ArXiv* 2020:03316. doi: 10.48550/arXiv.2007.03316

Han, X., Zhou, Y., Member, I. E. E. E., Chen, K., Qiu, H., Qiu, M., et al. (2023). ADS-Lead: lifelong anomaly detection in autonomous driving systems. *IEEE Trans. Intell. Transp. Syst.* 24:906. doi: 10.1109/TITS.2021.3122906

Hassija, V., Saxena, V., and Chamola, V. (2019). Scheduling drone charging for multi-drone network based on consensus time-stamp and game theory. *Comput. Commun.* 149, 51–61. doi: 10.1016/j.comcom.2019.09.021

He, Y., Chen, S., Wu, B., Xu, Y., and Xindong, W. (2021). *Unsupervised lifelong learning with curricula*. Proceedings of the Web Conference.

Hinton, G. E., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv* 2015:1503. doi: 10.48550/arXiv.1503.02531

Hwang, G. J., Xie, H., Wah, B. W., and Gašević, D. (2020). Vision, challenges, roles and research issues of artificial intelligence in education. *Comput. Educ.* 1:100001. doi: 10.1016/j.caeai.2020.100001

Irfan, B., Ramachandran, A., Spaulding, S., Kalkan, S., Parisi, G. I., and Gunes, H. (2021). Lifelong Learning and personalization in Long-term human-robot interaction(LEAP-HRI). HRI'21 Companion, Boulder, CO, USA.

Jan Moolman Buys University College University of Oxford. (2017). *Incremental generative models for syntactic and semantic natural language processing.*

Jeremy, F., et alet.al. (2013). *Autonomous Mission operations*. IEEE Aerospace Conference.

Jin, L., Liao, B., Liu, M., Xiao, L., Guo, D., and Yan, X. (2017). Different-level simultaneous minimization scheme for fault tolerance of redundant manipulator aided with discrete-time recurrent neural network. *Front. Neurorobot.* 11:50. doi: 10.3389/fnbot.2017.00050

Kahardipraja, P., Madureira, B., and Schlangen, D. (2023). TAPIR: learning adaptive revision for incremental natural language understanding with a two-Pass model. *arXiv* 2023:10845v1. doi: 10.48550/arXiv.2305.10845

Kahveci, N. E., and Ioannou, P. A. (2013). Adaptive steering control for uncertain ship dynamics and stability analysis. *Automatica* 49, 685–697. doi: 10.1016/j.automatica.2012.11.026

Kanero, J., Oranç, C., Koşkulu, S., Kumkale, G. T., Göksun, T., and Küntay, A. C. (2022). Are tutor robots for everyone? The influence of attitudes, anxiety, and personality on robot-led language learning. *Int. J. Soc. Robot.* 14, 297–312. doi: 10.1007/s12369-021-00789-3

Kang, M., Kang, M., and Han, B. (2022). *Class-incremental learning by knowledge distillation with adaptive feature consolidation*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR.

Katayama, H., and Aoki, H. (2014). Straight-line trajectory tracking control for sampled-data underactuated ships. *IEEE Trans. Control Syst. Technol.* 22, 1638–1645. doi: 10.1109/TCST.2013.2280717

Kemker, R., and Kanan, C. (2017). Fearnet: brain-inspired model for incremental learning. *arXiv* 2017:10563. doi: 10.48550/arXiv.1711.10563

Khan, M. A. (2022). Intelligent environment enabling autonomous driving. *Hindawi Comput. Intell. Neurosci.* 2022:2938011. doi: 10.1109/ACCESS.2021.3059652

Kim, D., and Han, B. (2023). On the stability-plasticity dilemma of class-incremental learning. *arXiv* 2023:01663v1. doi: 10.48550/arXiv.2304.01663

Kingston, D., Beard, R. W., and Holt, R. S. (2008). Decentralized perimeter surveillance using a team of UAVs. *IEEE Trans. Robot.* 24, 1394–1404. doi: 10.1109/TRO.2008.2007935

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci.* 114, 3521–3526. doi: 10.1073/pnas.1611835114

Klimova, B., Pikhart, M., Benites, A. D., Lehr, C., and Sanchez-Stockhammer, C. (2022). Neural machine translation in foreign language teaching and learning: a systematic review. *Educ. Inf. Technol.* 27, 1–20. doi: 10.1007/s10639-022-11194-2

Kober, J., Bagnell, J. A., and Peters, J. (2013). Reinforcement learning in robotics: a survey. *Int. J. Rob. Res.* 32, 1238–1274. doi: 10.1177/0278364913495721

Kou, X., Lin, Y., Liu, S., Li, P., Zhou, J., and Zhang, Y. (2020). *Disentangle-based continual graph representation learning*. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), pp. 2961–2972. Association for Computational Linguistics.

Kurz, G., Holoch, M., and Biber, P. (2021). *Geometry-based graph pruning for lifelong SLAM, 2021 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, Prague: Czech Republic, pp. 3313–3320.

Lesouple, J., Baudoin, C., Spigai, M., and Tourneret, J.-Y. (2021). Generalized isolation forest for anomaly detection. *Pattern Recogn. Lett.* 149, 109–119. doi: 10.1016/j.patrec.2021.05.022

Li, J., Dani, H., Hu, X., Tang, J., Chang, Y., and Liu, H. (2017). *Attributed network embedding for learning in a dynamic environment*. In: Proceedings of the 2017 ACM on conference on information and knowledge management (CIKM '17). New York: Association for Computing Machinery, pp. 387–396.

Li, Y., Qi, T., Ma, Z., Quan, D., and Miao, Q. (2023). Seeking a hierarchical prototype for multimodal gesture recognition. *IEEE Trans. Neural Netw. Learn. Syst.*:5811. doi: 10.1109/TNNLS.2023.3295811

Li, Z., Zhao, Y., Botta, N., Ionescu, C., and Hu, X. (2020). Copod: copulabased outlier detection. In: 2020 IEEE international conference on data mining (ICDM), IEEE, pp. 1118–1123.

Li, T. S., Zhao, R., Philip Chen, C. L., Fang, L. Y., and Liu, C. (2018). Finite-time formation control of under-actuated ships using nonlinear sliding mode control. *IEEE Trans. Cybern.* 48, 3243–3253. doi: 10.1109/TCYB.2018.2794968

Li, X., et al. (2023). *Graph structure-based implicit risk reasoning for Long-tail scenarios of automated driving*. 4th International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Hangzhou, China, pp. 415–420.

Li, Z., et al. (2023). *Two flexible translation-based models for knowledge graph embedding*. pp. 3093–3105.

Liao, B., Hua, C., Xu, Q., Cao, X., and Li, S. (2024). Inter-robot management via neighboring robot sensing and measurement using a zeroing neural dynamics approach. *Expert Syst. Appl.* 244:122938. doi: 10.1016/j.eswa.2023.122938

Lin, G., Chu, H., and Lai, H. (2022). *Towards better plasticity-stability trade-off in incremental learning: a simple linear connector*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Litman, T. (2021). *Autonomous vehicle implementation predictions: Implications for transport planning*. Victoria: Victoria Transport Policy Institute.

Liu, Z., Huang, C., Yu, Y., and Dong, J. (2021). Motif-preserving dynamic attributed network embedding. In: Proceedings of the web conference 2021 (WWW '21). New York: Association for Computing Machinery, 1629–1638.

Liu, H., Yang, Y., and Wang, X. (2021). Overcoming catastrophic forgetting in graph neural networks. *Proc. AAAI Conf. Artif. Intell.* 35, 8653–8661. doi: 10.1609/aaai.v35i10.17049

Lopez-Paz, D., and Ranzato, M. A. (2022). Gradient episodic memory for continual learning. *arXiv* 2022:08840v6. doi: 10.48550/arXiv.1706.08840

Lovinger, J., and Valova, I. (2020). Infinite lattice learner: an ensemble for incremental learning. *Soft. Comput.* 24, 6957–6974. doi: 10.1007/s00500-019-04330-7

Lv, X., Xiao, L., and Tan, Z. (2019). Improved Zhang neural network with finite-time convergence for time-varying linear system of equations solving. *Inf. Process. Lett.* 147, 88–93. doi: 10.1016/j.ipl.2019.03.012

Maguire, G., Ketz, N., Pilly, P. K., and Mouret, J. B. (2022). A-EMS: an adaptive emergency management system for autonomous agents in unforeseen situations. TAROS 2022: Towards Autonomous Robotic Systems, pp. 266–281.

Mallya, A., and Lazebnik, S. (2018a). PackNet: adding multiple tasks to a single network by iterative pruning. *arXiv* 2018:05769v2. doi: 10.48550/arXiv.1711.05769

Mallya, A., and Lazebnik, S. (2018b). *Packnet: adding multiple tasks to a single network by iterative pruning*. 2018 IEEE/CVF conference on computer vision and pattern recognition, Salt Lake City, UT, pp. 7765–7773.

May, R. D., and Loparo, K. A. (2014). *The use of software agents for autonomous control of a DC space power system*. AIAA-2014-3861, AIAA propulsion and energy forum, 12th international energy conversion engineering conference, Cleveland, OH.

May, R. D., et al. (2014). *An architecture to enable autonomous control of spacecraft*. AIAA-2014-3834, AIAA propulsion and energy forum, 12th international energy conversion engineering conference, Cleveland, OH.

Meng, X., Zhao, Y., Liang, Y., and Ma, X. (2022). Hyperspectral image classification based on class-incremental learning with knowledge distillation. *Remote Sens.* 14:2556. doi: 10.3390/rs14112556

Mohsan, S. A. H., Mohsan, S. A. H., Noor, F., Ullah, I., and Alsharif, M. H. (2022). Towards the unmanned aerial vehicles (UAVs): a comprehensive review. *Drones* 6:147. doi: 10.3390/drones6060147

Natali, A., Coutino, M., and Leus, G.. *Topology-aware joint graph filter and edge weight identification for network processes*. (2020). 2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP), Espoo, Finland, 2020, pp. 1–6.

Nguyen, T.-T., Pham, H. H., Le Nguyen, P., Nguyen, T. H., and Do, M. (2022). Multi-stream fusion for class incremental learning in pill image classification. *arXiv* 2022:02313v1. doi: 10.48550/arXiv.2210.02313

Nicolas, Y. (2018). Masse, Gregory D Grant, David J freeman. Alleviating catastrophic forgetting using context-based parameter modulation and synaptic stabilization. *Proc. Natl. Acad. Sci.* 115, E10467–E10475. doi: 10.1073/pnas.1803839115

Niemueller, T. (2013). *Stefan Schiffer, Gerhard Lakemeyer, Safoura Rezapour Lakani. Life-long Learning Perception using Cloud Database Technology*. Proceeding IROS Workshop on Cloud Robotics.

OM Group of Companies. (2020). *The future of automotive: still a Long* ride. Available at: www.omnicommediagroup.com/news/globalnews/the-future-of-automotive-st.

Ouyang, Y., Shi, J., Wei, H., and Gao, H. (2021). Incremental learning for personalized recommender systems. *arXiv* 2021:13299v1. doi: 10.48550/arXiv.2108.13299

Pan, S. J., and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22:191. doi: 10.1109/TKDE.2009.191

Pan, X., You, Y., Wang, Z., and Lu, C. (2017). *Virtual to real reinforcement learning for autonomous driving*. In: Proceedings of the British machine vision conference (BMVC). BMVA Press, London, UK.

Parisi, G. I., Tani, J., Weber, C., and Wermter, S. (2017). Lifelong learning of human actions with deep neural network self-organization. *Neural Netw.* 96, 137–149. doi: 10.1016/j.neunet.2017.09.001

Pikhart, M. (2021). Human-computer interaction in foreign language learning applications: applied linguistics viewpoint of mobile learning. *Proc. Comput. Sci.* 184, 92–98. doi: 10.1016/j.procs.2021.03.123

Qu, C., Zhang, L., Li, J., Deng, F., Tang, Y., Zeng, X., et al. (2021). Improving feature selection performance for classification of gene expression data using Harris hawks optimizer with variable neighborhood learning. *Brief. Bioinform.* 22:bbab097. doi: 10.1093/bib/bbab097

Rajan, V., and Cruz, A. D. L. (2022). Utilisation of service robots to assist human Workers in Completing Tasks Such in retail, hospitality, healthcare, and logistics businesses. *Technoarete Trans. Ind. Robot. Automat. Syst.* 2:2. doi: 10.36647/TTIRAS/02.01.A002

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., et al. (2022). Progressive neural networks. *arXiv* 2022:04671v4. doi: 10.48550/arXiv.1606.04671

Sarlin, P. E., DeTone, D., Malisiewicz, T., and Rabinovich, A. (2020). *SuperGlue: learning feature matching with graph neural networks*," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 4937–4946.

Schwarzerova, J., and Bajger, A. (2021). *Iro Pierdou, Lubos Popelinsky, Karel Sedlar, Wolfram Weckwerth. An innovative perspective on metabolomics data analysis in biomedical research using concept drift detection*. 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).

Serra, J., Suris, D., Miron, M., and Karatzoglou, A.. (2018). *Overcoming catastrophic forgetting with hard attention to the task*. 80. Proceedings of the 35th International Conference on Machine Learning, PMLR, pp. 4548–4557.

Shaheen, K., Hanif, M. A., Hasan, O., and Hasan, O. (2022). Continual learning for real-world autonomous systems: algorithms, challenges and frameworks. *J. Intell. Robot. Syst.* 105:9. doi: 10.48550/arXiv.2105.12374

She, Q., Feng, F., Liu, Q., Chan, R. H. M., Hao, X., Lan, C., et al. (2020). IROS 2019 lifelong robotic vision challenge lifelong object recognition report. *arXiv* 2020:14774v1. doi: 10.48550/arXiv.2004.14774

Shen, C., Shi, Y., and Buckham, B. (2019). Path-following control of an AUV: a multiobjective model predictive control approach. *IEEE Trans. Control Syst. Technol.* 27, 1334–1342. doi: 10.1109/TCST.2018.2789440

Shi, G., Chen, J., Zhang, W., Zhan, L.-M., and Xiao-Ming, W. (2021). Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima. *Adv. Neural Inf. Proces. Syst.* 34:21. doi: 10.48550/arXiv.2111.01549

Shi, X., Li, D., Zhao, P., Tian, Q., Tian, Y., Long, Q., et al. (2020). Are we ready for service robots? The OpenLORIS-scene datasets for lifelong SLAM. *arXiv* 2020:05603v2. doi: 10.48550/arXiv.1911.05603

Shin, H., Lee, J. K., Kim, J., and Kim, J. (2017). Continual learning with deep generative replay. *NIPS*, 2990–2999. doi: 10.48550/arXiv.1705.08690

Smith, J., Hsu, Y. C., Balloch, J., Shen, Y., Jin, H., and Kira, Z. (2021). *Always be dreaming: a new approach for data-free class-incremental learning*. 2021 IEEE/CVF International Conference on Computer Vison(ICCV).

Sodhani, S., Chandar, S., and Bengio, Y. (2019). Towards training recurrent neural networks for lifelong learning. *arXiv* 2019:07017v3. doi: 10.48550/arXiv.1811.07017

Soeder, J. F., Dever, T. P., McNelis, A. M., Beach, R. F., and Trase, L. M.. (2014). *Overview of intelligent power controller development for human deep space exploration*. 12th International Energy Conversion Engineering Conference, July 28–30.

Soeder, J. F., Dever, T. P., McNelis, A. M., Beach, R. F., Trase, L. M., and May, R. D. (2014). Overview of intelligent power controller development for human deep space exploration. AIAA-2014-3833, AIAA propulsion and energy forum, 12th international energy conversion engineering conference, Cleveland, OH.

Su, H., Qiu, M., and Wang, H. (2012). Secure wireless communication system for smart grid with rechargeable electric vehicles. *IEEE Commun. Mag.* 50, 62–68. doi: 10.1109/MCOM.2012.6257528

Su, J., Zou, D., Zhang, Z., and Wu, C. (2023). *Towards robust graph incremental learning on evolving graphs*. Proceedings of the 40th International Conference on Machine Learning Available at: https://proceedings.mlr.press/v202/su23a.html.

Sun, Q., and Wu, X. (2023). A deep learning-based approach for emotional analysis of sports dance. *PeerJ Comput. Sci.* 9:1441. doi: 10.7717/peerj-cs.1441

Tan, Z., Ding, K., Guo, R., and Liu, H. (2022). *Graph few-shot class-incremental learning*. In: Proceedings of the fifteenth ACM international conference on web search and data mining (WSDM '22). Association for Computing Machinery, New York, NY, USA, 987–996.

Tang, B., Zhong, Y., Neumann, U., Wang, G., Zhang, Y., and Chen, S. (2021). Collaborative uncertainty in multi-agent trajectory forecasting. In: *Proceedings of 35th conference on neural information processing systems (NeurIPS 2021)*.

Teixeira, K., Miguel, G., Silva, H. S., and Madeiro, F. (2023). A survey on applications of unmanned aerial vehicles using machine learning. *IEEE Access* 11, 117582–117621. doi: 10.1109/ACCESS.2023.3326101

Tian, H., Shuai, M., Li, K., and Peng, X. (2019). An incremental learning ensemble strategy for industrial process soft sensors. *Hindawi Complexity* 2019:5353296. doi: 10.1155/2019/5353296

Wang, W., Ni, H., Lin, S., Tao, H., Ren, Q., Gerstoft, P., et al. (2019). Deep transfer learning for source ranging: deep-sea experiment results. *J. Acoust. Soc. Am.* 146:EL317:EL322. doi: 10.1121/1.5126923

Wang, N., Qian, C., Sun, J.-C., and Liu, Y.-C. (2016). Adaptive robust finite-time trajectory tracking control of fully actuated marine surface vehicles. *IEEE Trans. Control Syst. Technol.* 24, 1454–1462. doi: 10.1109/TCST.2015.2496585

Wang, J., Song, G., Wu, Y., and Liang, W. (2020). *Streaming graph neural networks via continual learning*. In: proceedings of the 29th ACM international conference on information &amp; knowledge management (CIKM '20). Association for Computing Machinery, New York, USA, 1515–1524.

Wang, Y., Zhang, Y., and Coates, M. (2021). *Graph structure aware contrastive knowledge distillation for incremental learning in recommender systems*. CIKM '21: The 30th ACM International Conference on Information and Knowledge Management.

Wibisono, A., Piran, M. J., Song, H. K., and Lee, B. M. (2023). A survey on unmanned underwater vehicles: challenges, enabling technologies, and future research directions. *Sensors (Basel).* 23:7321. doi: 10.3390/s23177321

Wu, G., Gong, S., and Li, P. (2021). *Striking a balance between stability and plasticity for class-incremental learning*. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Xiang, Z., Xiang, C., Li, T., and Guo, Y. (2021). A self-adapting hierarchical actions and structures joint optimization framework for automatic design of robotic and animation skeletons. *Soft. Comput.* 25, 263–276. doi: 10.1007/s00500-020-05139-5

Xiao, Y. (2022). Application of machine learning in ethical Design of Autonomous Driving Crash Algorithms. *Hindawi Comput. Intell. Neurosci.* 2022:2938011. doi: 10.1155/2022/2938011

Yan, S., Xie, J., and He, X. (2021). DER: dynamically expandable representation for class incremental learning. *arXiv* 2021:16788v1. doi: 10.48550/arXiv.2103.16788

Yang, Q. (2013). *Big data, lifelong machine learning and transfer learning*. Proceedings of the Sixth ACM International Conference on Web Search and Data Mining.

Yoon, H. G., Song, H. J., Park, S. B., and Park, S. Y. (2016). *A translation-based knowledge graph embedding preserving logical property of relations*. In: Proceedings of the 2016 conference of the north American chapter of the Association for Computational Linguistics: Human language technologies, pp. 907–916.

Yoon, J., Yang, E., Lee, J., and Hwang, S. J. (2017). Lifelong learning with dynamically expandable networks. *ArXiv* 2017:01547. doi: 10.48550/arXiv.1708.01547

Zenke, F., Poole, B., and Ganguli, S. (2017a). *Continual learning through synaptic intelligence*. Proceedings of the 34th International Conference on Machine Learning, No. 70, pp. 3987–3995.

Zenke, F., Poole, B., and Ganguli, S. (2017b). Continual learning through synaptic intelligence. *arXiv* 2017:04200v3. doi: 10.48550/arXiv.1703.04200

Zhang, X., Song, D., and Tao, D. (2023a). Hierarchical prototype networks for continual graph representation learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 4622–4636. doi: 10.1109/TPAMI.2022.3186909

Zhang, X., Song, D., and Tao, D. (2023b). Ricci curvature-based graph Sparsification for continual graph representation learning. *IEEE Trans. Neural. Netw. Learn Syst.*:454. doi: 10.1109/TNNLS.2023.3303454

Zhang, J., Wang, T., Ng, W. W. Y., and Pedrycz, W. (2023). KNNENS: a k-nearest neighbor ensemble-based method for incremental learning under data stream with emerging new classes. *IEEE Trans. Neural Netw. Learn. Syst.* 34, 9520–9527. doi: 10.1109/TNNLS.2022.3149991

Zhang, Q., et al. (2022). *A dynamic Variational framework for open-world node classification in structured sequences*. 2022 IEEE International Conference on Data Mining (ICDM), Orlando, FL, USA, pp. 703–712.

Zhao, Y., Lin, H., Wang, H., and Ma, X. (2022). *Few-SHOT class incremental learning for HYPERSPECTRAL image classification based on constantly updated CLASSIFIER. GARSS* 2022–2022 IEEE International Geoscience and Remote Sensing Symposium.

Zhao, C., Song, A., Zhu, Y., Jiang, S., Liao, F., Yuchuan, D., et al. (2023). Data-driven indoor positioning correction for infrastructure-enabled autonomous driving systems: a lifelong framework. *IEEE Trans. Intell. Transp. Syst.* 24:563. doi: 10.1109/TITS.2022.3233563

Zhao, H., Wang, H., Yongjian, F., Wu, F., and Li, X. (2021). Memory efficient class-incremental learning for image classification. *IEEE Trans. Neural Net. Learn. Syst.* 11:1. doi: 10.48550/arXiv.2008.01411

Zhou, F., and Cao, C. (2021a) Overcoming catastrophic forgetting in graph neural networks with experience replay. *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21).*

Zhou, F., and Cao, C. (2021b). Overcoming catastrophic forgetting in graph neural networks with experience replay. *Proc. AAAI Conf. Artif. Intell.* 35, 4714–4722. doi: 10.1609/aaai.v35i5.16602

Zhou, W., Cao, Z., Xu, Y., Deng, N., Liu, X., Jiang, K., et al. (2022). *Long-tail prediction uncertainty aware trajectory planning for self-driving vehicles*. 2022 IEEE 25th international conference on intelligent transportation systems (ITSC), Macau, China, pp. 1275–1282.

Zhou, W., Chen, D., Yan, J., Li, Z., Yin, H., and Ge, W. (2022). Multi-agent reinforcement learning for cooperative lane changing of connected and autonomous vehicles in mixed traffic. *Auton. Intell. Sys.* 1, 2–5. doi: 10.48550/arXiv.2111.06318

Zhou, D.-W., Ye, H.-J., Liang, M., Xie, D., Shiliang, P., and Zhan, D.-C. (2022). Few-shot class-incremental learning by sampling multi-phase tasks. *arXiv* 2022:17030v2. doi: 10.48550/arXiv.2203.17030

Zhou, Y., Zhang, X., Wang, S., and Li, L.. (2022) *Multi-attribute joint point cloud super-resolution with adversarial feature graph networks*. 2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Taipei City, Taiwan, 2022: pp. 1–6.

Zhu, X. L., Wang, H. C., You, H. M., Zhang, W. H., Zhang, Y. Y., Liu, S., et al. (2021). Survey on testing of intelligent Systems in Autonomous Vehicles. *J. Softw.* 32, 2056–2077. doi: 10.13328/j.cnki.jos.006266

Zhu, F., Zhang, X. Y., Wang, C., Yin, F., and Liu, C. L. (2021). *Prototype Augmentation and Self-Supervision for Incremental Learning.*

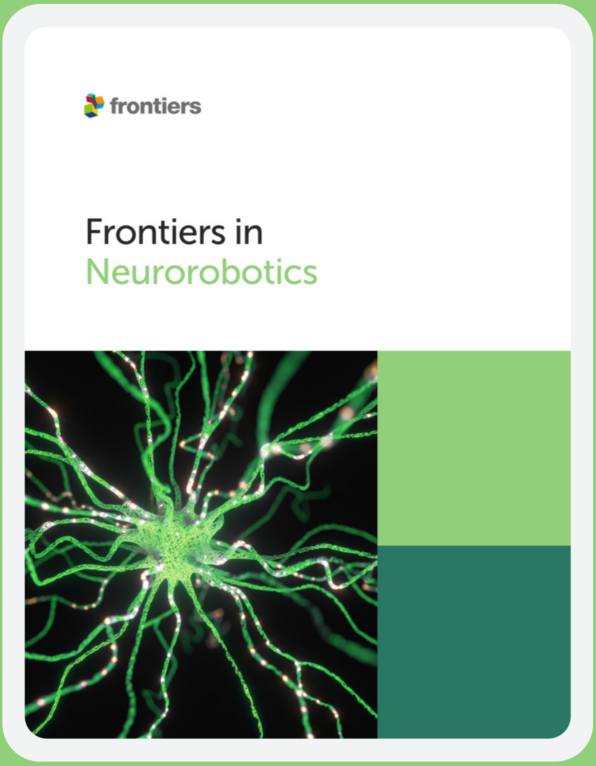
Frontiers in Neurorobotics

Investigates embodied autonomous neural systems and their impact on our lives

Part of the most cited neuroscience series, this journal advances understanding of neurorobotics - from prosthetic devices to brain machine interfaces, and wearable systems to home appliances.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact