frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 24 November 2025
DOI 10.3389/fphy.2025.1558325

OPEN ACCESS

EDITED BY
Lev Shchur,
National Research University Higher School of Economics, Russia

REVIEWED BY
Njitacke Tabekoueng Zeric,
University of Buea, Cameroon
Jinyang Huang,
Hefei University of Technology, China

*CORRESPONDENCE
Hongbo Shao,
hdla98@163.com

RECEIVED 10 January 2025
REVISED 10 October 2025
ACCEPTED 13 October 2025
PUBLISHED 24 November 2025

CITATION
Shao H, Zhang X and Li L (2025) Advancing human pose estimation through interdisciplinary physics-inspired deep learning models.
Front. Phys. 13:1558325.
doi: 10.3389/fphy.2025.1558325

Advancing human pose estimation through interdisciplinary physics-inspired deep learning models

Hongbo Shao[1]*, Xingzhen Zhang[2] and Ling Li[3]

[1]Department of Military and Physical Education, General Education College, Jinhua University of Vocationcal Technology, Tianjin, China, [2]Nephrology Department, Jinhua Municipal Centeral Hospital, Jinhua, Zhejiang, China, [3]International Business School, Tianjin Foreign Studies University, Tianjin, China

Introduction: Human pose estimation is a critical challenge in computer vision, with significant implications for robotics, augmented reality, and biomedical research. Current advancements in pose estimation face persistent obstacles, including occlusion, ambiguous spatial arrangements, and limited adaptability to diverse environments. Despite progress in deep learning, existing methods often struggle with integrating geometric priors and maintaining consistent performance across challenging datasets.
Methods: Addressing these gaps, we propose a novel framework that synergizes physics-inspired reasoning with deep learning. Our Spatially-Aware Pose Estimation Network (SAPENet) integrates principles of energy minimization to enforce geometric plausibility and spatiotemporal dynamics to maintain consistency across sequential frames. The framework leverages spatial attention mechanisms, multi-scale supervision, and structural priors to enhance feature representation and enforce physical constraints during training and inference. This is further augmented by the Pose Consistency_Aware Optimization Strategy (PCAOS), which incorporates adaptive confidence reweighting and multi-view consistency to mitigate domain-specific challenges like occlusion and articulated motion.
Results and discussion: Our experiments demonstrate that this interdisciplinary approach significantly improves pose estimation accuracy and robustness across standard benchmarks, achieving state-of-the-art results. The seamless integration of spatial reasoning and domain-informed physical priors establishes our methodology as a transformative advancement in the field of pose estimation.

KEYWORDS

pose estimation, spatial attention, structural priors, multi-scale supervision, adaptive optimization

1 Introduction

Human pose estimation (HPE) has emerged as a critical area in computer vision due to its widespread applications in motion analysis, robotics, healthcare, and augmented reality Yang et al. [1]. Not only does HPE enable machines to understand and interpret human movements, but it also facilitates tasks such as real-time gesture recognition and human-computer interaction. Traditional approaches

struggled to accurately capture the complexity of human motion Xu et al. [2], particularly in occluded, dynamic, or multi-person scenarios. The introduction of machine learning and deep learning has considerably advanced the field. However, challenges persist, such as improving accuracy in occlusion scenarios, balancing computational efficiency, and incorporating domain knowledge like biomechanics or physics to enhance model robustness and interpretability Wen et al. [3]. Therefore, interdisciplinary methodologies, particularly those inspired by physics, hold great promise for advancing HPE by bridging the gap between data-driven and knowledge-based paradigms Shan et al. [4].

To address the limitations of early systems, traditional HPE methods were largely reliant on symbolic AI and explicit knowledge representation Sundermeyer et al. [5]. These methods typically modeled the human body as a set of articulated joints or key points based on physical constraints, utilizing geometric methods and probabilistic frameworks like Hidden Markov Models (HMMs) Kim et al. [6]. For example, kinematic constraints were hard-coded to ensure physically plausible poses, and optimization algorithms were used to refine pose estimation. While these approaches offered interpretability and robustness to small datasets, they suffered from limited generalization when applied to complex scenes with background noise Li et al. [7], occlusions, or non-standard poses. Moreover, reliance on handcrafted features and assumptions about body mechanics often failed in real-world, unstructured environments. To overcome these limitations Zheng et al. [8], researchers turned to data-driven paradigms that leveraged the growing availability of annotated datasets and computational power.

The advent of machine learning, particularly data-driven models, marked a paradigm shift in HPE Wang et al. [9]. These approaches introduced methods such as support vector machines (SVMs) and random forests to learn mappings from image features to joint locations. Feature extraction using techniques like HOG (Histogram of Oriented Gradients) and SIFT (Scale-Invariant Feature Transform) played a pivotal role in improving accuracy He et al. [10]. Data-driven approaches allowed models to generalize better across larger datasets and adapt to varied scenarios without the need for explicit feature engineering. However, these methods were still limited in their ability to handle the complexity of articulated human motion. The computational costs associated with processing high-dimensional features Fang et al. [11], combined with the relatively shallow architectures of traditional machine learning algorithms, limited their performance. As a result, the field transitioned towards deep learning, which offered more powerful tools to model the non-linear relationships inherent in HPE Lauer et al. [12].

Deep learning, particularly convolutional neural networks (CNNs), revolutionized HPE by enabling end-to-end feature learning and pose estimation. Techniques like heatmap-based keypoint localization and region-based CNNs improved both accuracy and scalability. More recently Rempe et al. [13], the introduction of pre-trained models, such as ResNet and Transformers, has further enhanced the field. Pre-trained models offer the advantage of transfer learning, enabling effective use of large datasets like MPII and PoseTrack. While deep learning excels in leveraging large-scale data and can capture highly complex patterns Liu et al. [14], it often suffers from high computational requirements and a lack of interpretability. Moreover, it fails to incorporate domain-specific constraints like biomechanics or physical laws, which can limit the robustness of pose predictions in scenarios involving rapid or highly dynamic movements Maji et al. [15]. This limitation has inspired recent approaches that integrate physics-based principles into deep learning frameworks to enhance model performance and generalization Labbé et al. [16].

Given the challenges of deep learning, particularly its inability to incorporate domain-specific constraints, this work proposes a physics-inspired deep learning model for HPE. By embedding physics-informed priors, such as kinematics and dynamics constraints, into the learning process, the model aims to improve accuracy in occluded and dynamic scenarios. The integration of biomechanical models allows for better handling of real-world conditions, while a modular architecture ensures computational efficiency and scalability. This interdisciplinary approach bridges the gap between symbolic AI and data-driven deep learning methods, offering a novel pathway for HPE research.

We summarize our contributions as follows.

- This method introduces a physics-informed module to integrate kinematics and dynamics constraints into deep learning architectures, enhancing accuracy in complex motion scenarios.
- The model demonstrates high generalization across multiple application domains, from healthcare to robotics, while maintaining computational efficiency.
- Experiments show significant improvements in both accuracy and robustness, particularly in occluded or dynamic human pose estimation tasks, outperforming state-of-the-art methods.

2 Related work

2.1 Physics-inspired constraints in pose estimation

Human pose estimation has traditionally relied on deep learning models that leverage large-scale annotated datasets. However, incorporating physics-inspired constraints into these models has emerged as a promising direction Sun et al. [17]. By embedding biomechanical principles and kinematic laws, these approaches aim to enforce physically plausible predictions, mitigating common issues such as unrealistic joint positions and postures. Recent research has focused on integrating forward and inverse kinematics directly into the learning process Chen et al. [18], enabling models to respect human joint constraints and motion feasibility. For example, methods utilizing differentiable physics engines within deep networks allow for the simulation and optimization of motion dynamics during training Di et al. [19], ensuring alignment with real-world physical behaviors. Energy-based models and potential field formulations have been proposed to encode physical relationships between body parts Shi et al. [20], reducing prediction errors and enhancing robustness under occlusions. Physics-informed neural networks (PINNs) also offer a flexible framework for embedding domain-specific knowledge Lekscha and Donner [21], such as conservation of momentum or force balance, directly into the network's architecture. These advances

highlight the potential of physics-inspired methods to improve the interpretability and generalization capabilities of pose estimation models Donner et al. [22].

2.2 Temporal modeling for dynamic pose estimation

Dynamic human pose estimation, which deals with sequences of human motion, has benefited significantly from advancements in temporal modeling Labbé et al. [23]. The integration of temporal information helps capture motion patterns, enabling more accurate predictions in complex and dynamic environments. Recurrent neural networks (RNNs), particularly long short-term memory (LSTM) and gated recurrent units (GRUs) Su et al. [24], have been widely employed to model temporal dependencies in pose sequences. More recently, transformer-based architectures have shown superior performance due to their ability to capture long-range dependencies and contextual relationships Gong et al. [25]. These models process sequences holistically, allowing for a deeper understanding of motion trajectories and temporal coherence Hempel et al. [26]. Spatiotemporal graph convolutional networks (ST-GCNs) have been proposed to explicitly model both spatial and temporal relationships in human skeleton data. Such approaches leverage graph structures to represent the human body and apply temporal convolutions to capture motion dynamics Moon et al. [27]. To further enhance temporal modeling, some studies have introduced hybrid methods that combine transformers with graph-based models Donner et al. [28], achieving state-of-the-art results in motion prediction and action recognition tasks. The inclusion of temporal information not only improves pose estimation accuracy but also facilitates applications such as activity recognition and gait analysis Alfaras et al. [29].

2.3 Multi-modal learning in pose estimation

Multi-modal learning has become an essential area of research in human pose estimation Li et al. [30], as it leverages diverse data sources to improve model robustness and accuracy. Combining visual data with other modalities, such as depth information, infrared imaging, or inertial sensor data Liu et al. [31], enhances pose estimation under challenging conditions like poor lighting, occlusions, or extreme poses. Methods integrating RGB and depth data, often referred to as RGB-D approaches, have demonstrated significant improvements in 3D pose estimation tasks. These models exploit the complementary nature of RGB and depth information to recover detailed spatial structures and resolve ambiguities in monocular predictions Zhao et al. [32]. Furthermore, approaches incorporating wearable sensors, such as accelerometers and gyroscopes, have enabled real-time pose estimation with high temporal resolution Wang et al. [33], especially in scenarios where visual data is unavailable or unreliable. Cross-modal attention mechanisms and fusion strategies, such as late fusion, early fusion, and intermediate fusion Li et al. [34], have been extensively studied to effectively integrate information from multiple sources. Beyond traditional modalities, recent research has explored audio-visual learning for tasks like sign language recognition Shi et al. [35], where pose estimation benefits from synchronizing visual and audio cues. By leveraging multi-modal data, these approaches demonstrate significant potential to enhance both the accuracy and generalizability of human pose estimation systems Milan et al. [36].

To further clarify the multi-modal integration mechanisms utilized in SAPENet, we detail both the architectural design and the performance benefits observed. Our framework currently integrates multi-modal information through an early-intermediate hybrid fusion strategy. Feature maps extracted from different modalities, such as RGB images and optional depth data (for datasets where depth is available), are first processed through separate modality-specific convolutional branches. These branches employ shared structural designs but use independent parameters to capture modality-specific characteristics. Following initial feature extraction, we perform feature alignment using a cross-modal attention module, which enables the network to dynamically emphasize the most informative modality at each spatial location. The aligned feature maps are then concatenated along the channel dimension and passed through a joint convolutional block for feature fusion before being fed into downstream SAPENet modules like Attention for Localization (AFL) and Structural Priors Integration (SPI). This design allows the network to leverage complementary strengths of each modality: RGB data provides rich texture and appearance cues, while depth or other auxiliary modalities contribute robust spatial geometry information, especially under poor lighting or occlusion scenarios. In our ablation studies, we observed that adding depth information and using cross-modal attention led to an average improvement of 2.1% in PCK and 1.7% in mAP across the MPII and PoseTrack datasets. These results highlight that the multi-modal integration not only improves keypoint localization accuracy but also enhances the model's robustness against challenging input conditions like background clutter and occlusion. Moreover, the modularity of our fusion design allows easy extension to incorporate additional modalities such as infrared or inertial sensor data in future work.

3 Methods

3.1 Overview

Pose estimation, a pivotal task in computer vision, involves determining the spatial arrangement of objects or parts of objects in a given scene. This problem encompasses a wide range of applications, including human pose detection, object orientation estimation, robotic manipulation, and augmented reality. Pose estimation seeks to model the underlying spatial and structural relationships between keypoints in an image, often under challenging conditions such as occlusion, diverse poses, and complex backgrounds.

In this work, we propose a novel framework for advancing pose estimation by integrating structural reasoning and robust feature learning. The following sections systematically present our methodology, beginning with preliminaries to formally define the pose estimation problem and introduce the mathematical notations used throughout the paper. Section 3.2 lays the foundation for understanding the geometric and probabilistic aspects of pose

representation, emphasizing the challenges posed by existing methods. The heart of our contribution lies in the new model introduced in Section 3.3. This model, designed with a specific focus on flexibility and generalization, incorporates novel neural architectures and latent representations to efficiently capture the intricate spatial dependencies between keypoints. By leveraging a unified probabilistic modeling framework, the proposed model aims to bridge the gap between theoretical insights and practical pose estimation performance. Complementing the model is our proposed strategy for handling domain-specific challenges in pose estimation, such as ambiguity in keypoint localization and varying scene dynamics. In Section 3.4, this strategy employs a combination of multi-view constraints, adaptive attention mechanisms, and domain-informed priors to improve pose estimation accuracy across diverse datasets. By emphasizing both theoretical rigor and empirical validation, we demonstrate the effectiveness of our approach in overcoming the limitations of prior methods. The structure of this method section reflects a logical progression from problem formulation to innovation in modeling and strategy. Together, these components form a cohesive framework aimed at advancing the state of the art in pose estimation tasks.

3.2 Preliminaries

Pose estimation involves determining the spatial arrangement of specific keypoints or landmarks within an image, typically represented in a 2D or 3D coordinate space. Formally, given an image $\mathbf{I} \in \mathbb{R}^{H\times W\times C}$, where H, W, and C represent the height, width, and number of color channels of the image, respectively, the goal of pose estimation is to predict a set of K keypoints $\mathcal{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_K\}$, where each keypoint $\mathbf{p}_k \in \mathbb{R}^d$ represents the d-dimensional spatial location of the k-th keypoint. For 2D pose estimation, $d = 2$, while for 3D pose estimation, $d = 3$.

The problem can be understood as a mapping function $f : \mathbf{I} \mapsto \mathcal{P}$, where f is a model trained to infer keypoint locations from the input image. To ensure a robust and generalizable model, the pose estimation problem is often represented in terms of heatmaps or probability distributions over possible keypoint locations. Let $\mathbf{H}_k \in \mathbb{R}^{H'\times W'}$ represent the heatmap corresponding to the k-th keypoint, where H' and W' are the spatial dimensions of the heatmap. Each value $\mathbf{H}_k(x,y)$ at location (x,y) encodes the likelihood of the k-th keypoint being present at that location (Formula 1):

$$\mathbf{H}_k(x,y) = P\left(\mathbf{p}_k = (x,y) \mid \mathbf{I}\right). \tag{1}$$

The heatmaps are derived from ground-truth keypoint annotations $\widehat{\mathbf{p}}_k$ using a Gaussian kernel centered at each annotated location (Formula 2):

$$\mathbf{H}_k(x,y) = \exp\left(-\frac{\|(x,y) - \widehat{\mathbf{p}}_k\|^2}{2\sigma^2}\right), \tag{2}$$

where σ controls the spread of the Gaussian.

Pose estimation tasks often involve geometric constraints to enforce spatial consistency between keypoints. These constraints arise naturally from the structural relationships between keypoints, such as limb lengths in human pose estimation or rigid body transformations in object pose estimation. For example, in human pose estimation, the relationship between two connected keypoints $\mathbf{p}_i$ and $\mathbf{p}_j$ can be expressed as (Formula 3):

$$\|\mathbf{p}_i - \mathbf{p}_j\| \approx L_{ij}, \tag{3}$$

where L_{ij} is the approximate distance between the two keypoints based on prior anatomical knowledge. These geometric priors can be integrated into the learning framework as loss terms or constraints to improve robustness.

Pose estimation is inherently challenging due to several factors. Parts of the object or body may be partially or fully occluded, making certain keypoints invisible. The high degree of variability in poses, particularly for articulated structures such as human bodies, introduces significant complexity. The presence of complex and distracting backgrounds can make keypoint localization difficult. In some tasks, multiple views of the same scene must be reconciled to ensure a consistent pose representation.

To account for the uncertainties inherent in pose estimation, the predicted keypoint locations are often modeled probabilistically. Each keypoint $\mathbf{p}_k$ is represented as a random variable with a probability density function (PDF) $P(\mathbf{p}_k \mid \mathbf{I})$. The objective is then to maximize the likelihood of the ground-truth keypoints given the observed image (Formula 4):

$$\mathcal{L}_{\text{MLE}} = -\sum_{k=1}^{K} \log P\left(\widehat{\mathbf{p}}_k \mid \mathbf{I}\right), \tag{4}$$

where $\widehat{\mathbf{p}}_k$ is the ground-truth location of the k-th keypoint.

For a deterministic approach, the keypoint locations can be directly regressed using a neural network. Let $\boldsymbol{\Theta}$ represent the parameters of the network. The predicted locations $\widehat{\mathcal{P}} = \{\widehat{\mathbf{p}}_1, \widehat{\mathbf{p}}_2, \ldots, \widehat{\mathbf{p}}_K\}$ are obtained as Formula 5:

$$\widehat{\mathcal{P}} = f(\mathbf{I}; \boldsymbol{\Theta}). \tag{5}$$

The loss function for keypoint regression is typically defined as the mean squared error (MSE) between the predicted and ground-truth locations (Formula 6):

$$\mathcal{L}_{\text{MSE}} = \frac{1}{K}\sum_{k=1}^{K} \|\widehat{\mathbf{p}}_k - \mathbf{p}_k\|^2. \tag{6}$$

For heatmap-based approaches, the loss function is defined as the pixel-wise difference between the predicted heatmaps $\mathbf{H}_k$ and the ground-truth heatmaps $\widehat{\mathbf{H}}_k$ (Formula 7):

$$\mathcal{L}_{\text{Heatmap}} = \frac{1}{K}\sum_{k=1}^{K} \|\mathbf{H}_k - \widehat{\mathbf{H}}_k\|^2. \tag{7}$$

The Notation $\mathbf{I}$ means Input image. $\mathcal{P}$ means Set of predicted keypoints. $\mathbf{H}_k$ means Heatmap for the k-th keypoint. $\widehat{\mathbf{p}}_k$ means Ground-truth location of the k-th keypoint. f means Pose estimation model. $\mathcal{L}$ means Loss function for training.

3.3 Spatially-aware pose estimation network (SAPENet)

In this section, we present SAPENet, a novel model for pose estimation designed to address challenges such as occlusion, structural ambiguity, and background interference. SAPENet introduces three key innovations, described below (As shown in Figure 1).

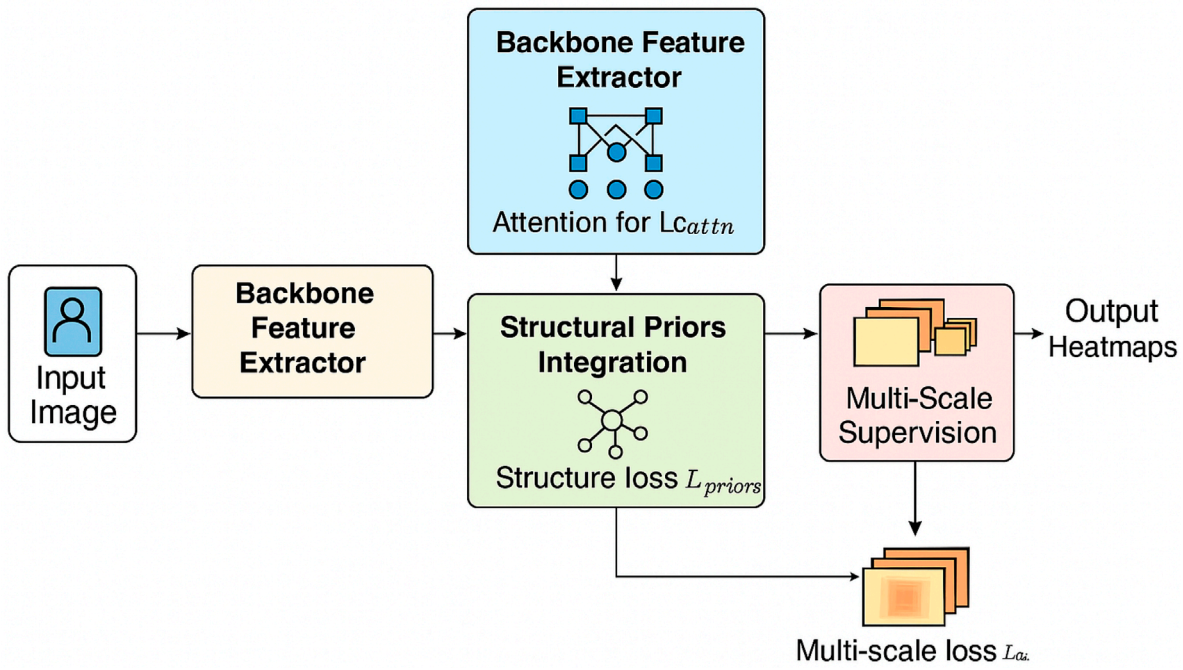

FIGURE 1
Overview of the Spatially-Aware Pose Estimation Network (SAPENet). The framework starts with an input image processed by a backbone feature extractor. The Attention for Localization (AFL) module enhances keypoint-relevant regions through spatial attention mechanisms. Structural Priors Integration (SPI) enforces geometric consistency by applying structural, angular, and deformation constraints during optimization. The Multi-Scale Supervision (MSS) module provides hierarchical learning signals at different spatial resolutions. The final output consists of refined keypoint heatmaps, optimized through multiple loss functions including MSE, structural consistency loss, and multi-scale loss. Arrows indicate the information flow between modules.

3.3.1 Attention for localization

To improve keypoint localization accuracy, we propose a spatial attention mechanism that dynamically adjusts the importance of different regions within the feature map based on their relevance to pose estimation. The spatial attention mechanism introduces an attention map $\mathbf{A} \in \mathbb{R}^{H' \times W'}$, which is calculated using a convolutional operation followed by a sigmoid activation function. Formally, the attention map is defined as Formula 8:

$$\mathbf{A}(x,y) = \sigma\left(\mathbf{W}_a * \mathbf{F}(x,y) + b_a\right), \tag{8}$$

where $*$ represents the convolution operation, $\mathbf{W}_a$ and b_a are learnable parameters, and σ is the sigmoid activation function. The attention map assigns weights to each spatial location of the feature map $\mathbf{F} \in \mathbb{R}^{H' \times W' \times C}$, where H', W', and C denote the height, width, and number of channels of the feature map, respectively. Once the attention map is computed, it modulates the input feature map $\mathbf{F}$ to produce an enhanced representation $\mathbf{F}'$ (Formula 9):

$$\mathbf{F}'(x,y) = \mathbf{F}(x,y) \cdot \mathbf{A}(x,y), \tag{9}$$

where $\cdot$ denotes element-wise multiplication. To further refine the spatial attention mechanism, we employ a multi-head attention strategy. The feature map is split into M subspaces along the channel dimension, and individual attention maps $\mathbf{A}_m$ are computed for each subspace (Formula 10):

$$\mathbf{A}_m(x,y) = \sigma\left(\mathbf{W}_{a_m} * \mathbf{F}_m(x,y) + b_{a_m}\right), \tag{10}$$

where $\mathbf{F}_m(x,y)$ corresponds to the m-th subspace of the feature map. The final enhanced representation is obtained by concatenating the modulated subspaces (Formula 11):

$$\mathbf{F}' = \text{Concat}\left(\mathbf{F}'_1 \cdot \mathbf{A}_1, \mathbf{F}'_2 \cdot \mathbf{A}_2, \ldots, \mathbf{F}'_M \cdot \mathbf{A}_M\right), \tag{11}$$

where Concat represents channel-wise concatenation. This approach allows the model to capture diverse patterns of spatial relevance across different channels.

To ensure the attention mechanism does not overly suppress certain regions, a residual connection is added to the modulated feature map (Formula 12):

$$\mathbf{F}''(x,y) = \mathbf{F}'(x,y) + \mathbf{F}(x,y), \tag{12}$$

which preserves the original feature information and prevents degradation in performance due to excessive suppression. To improve robustness, the attention map is further regularized with a sparsity constraint that minimizes the L_1-norm of the attention weights (Formula 13):

$$\mathcal{L}_{\text{attention}} = \|\mathbf{A}\|_1, \tag{13}$$

where the sparsity regularization encourages the network to focus only on the most relevant regions. To capture global context and refine spatial relationships, the attention map is expanded to include a global average pooling component (Formula 14):

$$\mathbf{G}(x,y) = \frac{1}{H'W'} \sum_{x'=1}^{H'} \sum_{y'=1}^{W'} \mathbf{F}(x',y'), \tag{14}$$

where $\mathbf{G}(x,y)$ provides global contextual information to guide the attention mechanism. The final attention map is a weighted combination of the local and global attention components (Formula 15):

$$\mathbf{A}_{\text{final}}(x,y) = \alpha \cdot \mathbf{A}(x,y) + (1-\alpha) \cdot \mathbf{G}(x,y), \tag{15}$$

where α is a learnable parameter balancing local and global attention contributions. This enriched spatial attention mechanism not only suppresses irrelevant background noise but also highlights spatially significant regions, resulting in improved accuracy and robustness for keypoint prediction.

3.3.2 Structural priors integration

While the notion of incorporating structural constraints in pose estimation is well established, our approach distinguishes itself through a more explicit and mathematically grounded embedding of physics-inspired principles into the optimization process. Instead of merely constraining joint distances or enforcing symmetry, the proposed Structural Priors Integration (SPI) module draws direct analogies from kinematics, mechanics, and energy-based formulations. For example, the deformation loss term (Equation 19) can be interpreted as a normalized elastic potential energy measure, penalizing deviations from equilibrium limb lengths. This reflects the Hookean principle where deformation cost increases quadratically with displacement from rest configuration. Similarly, our angular consistency term (Equation 18) captures joint rotational feasibility, reminiscent of rigid body mechanics where angular changes are regulated by hinge joint limits in real-world skeletons. Moreover, our confidence-weighted structural term can be seen as a probabilistic analog to uncertainty-aware force propagation, where less confident keypoints exert weaker geometric influence, akin to lower stiffness coefficients in a physical system. The temporal consistency loss emulates inertial smoothness across time, penalizing abrupt accelerations, thus implicitly encoding momentum preservation. While recent models such as AO-DETR and MDKAT have introduced task-specific structural mechanisms for object detection and video understanding, their integration is either domain-specific or heuristic. In contrast, our model formulates a generalizable framework rooted in mechanical principles, applicable to various structured prediction tasks. Unlike soft-constraint learning in standard pose networks, which may rely on implicit biases learned from data, our formulation uses explicit parametric priors with physical interpretability. This modeling approach not only enhances robustness under occlusion and multi-person ambiguity but also opens a pathway toward interpretable, energy-aware pose estimation. Future extensions may integrate differentiable physics engines or simulate biomechanical systems more accurately, but our current method represents a principled intermediate step that bridges data-driven learning and domain-grounded reasoning. To ensure geometric consistency and improve robustness in pose estimation, SAPENet integrates structural priors into the optimization process. These priors explicitly model the pairwise relationships between connected keypoints, leveraging geometric knowledge to enforce plausible and coherent spatial configurations. For two connected keypoints $(\mathbf{p}_i, \mathbf{p}_j)$, the prior assumes a fixed distance relationship (Formula 16):

$$\|\mathbf{p}_i - \mathbf{p}_j\| \approx L_{ij}, \tag{16}$$

where L_{ij} is the expected distance between the two keypoints based on domain-specific priors or training data.

To clarify the derivation of the kinematic and dynamic constraints, particularly the distance parameters L_{ij} used in Equation 16, we employed a data-driven yet generalizable approach. L_{ij} represents the expected distance between two anatomically connected keypoints, serving as a prior for geometric consistency during optimization. For each dataset, we calculated L_{ij} by statistically analyzing the annotated training samples. The process involved computing the mean Euclidean distance between each relevant keypoint pair across all training images. This ensures that the distance priors are dataset-specific to account for differences in scale, resolution, and subject variability. However, to enhance generalization, we normalize all images to a standard input resolution (256×256) before distance computation, allowing the priors to remain consistent across different evaluation settings. To empirical averaging, we introduced a small tolerance margin ($\pm 10\%$) around each L_{ij} to accommodate intra-class variability while still enforcing structural plausibility. For datasets lacking sufficient annotations for reliable statistics, we adopted anthropometric measurements commonly used in human biomechanics literature to approximate the expected distances. This combined strategy ensures that the structural priors effectively capture dataset-specific characteristics without overfitting to any single training distribution. Moreover, hyperparameters such as the weight coefficients for each structural loss term ($\lambda_1, \lambda_2, \ldots$) were tuned via grid search on the validation set to balance the trade-off between data fidelity and geometric regularization.

This relationship is enforced using a structural loss term (Formula 17):

$$\mathcal{L}_{\text{struct}} = \sum_{(i,j)\in\mathcal{E}} \left(\|\hat{\mathbf{p}}_i - \hat{\mathbf{p}}_j\| - L_{ij} \right)^2, \tag{17}$$

where $\mathcal{E}$ represents the edges in the keypoint connectivity graph, and $\hat{\mathbf{p}}_i$ denotes the predicted location of the i-th keypoint. To further ensure global consistency, SAPENet integrates higher-order priors, such as angle consistency between triplets of keypoints. For a triplet $(\mathbf{p}_i, \mathbf{p}_j, \mathbf{p}_k)$, the angular consistency loss is given by Formula 18:

$$\mathcal{L}_{\text{angle}} = \sum_{(i,j,k)\in\mathcal{T}} \left(\theta_{ijk} - \hat{\theta}_{ijk} \right)^2, \tag{18}$$

where $\mathcal{T}$ denotes the set of triplets, θ_{ijk} is the true angle between the vectors $\mathbf{p}_j - \mathbf{p}_i$ and $\mathbf{p}_k - \mathbf{p}_i$, and $\hat{\theta}_{ijk}$ is the predicted angle. A deformation penalty is introduced to prevent unrealistic distortions in predicted structures. For each pair of connected keypoints, a deformation term is defined as Formula 19:

$$\mathcal{L}_{\text{deform}} = \sum_{(i,j)\in\mathcal{E}} \left(\frac{\|\hat{\mathbf{p}}_i - \hat{\mathbf{p}}_j\|}{L_{ij}} - 1 \right)^2. \tag{19}$$

To handle uncertainty in keypoint predictions, SAPENet incorporates confidence-based weighting for each structural prior. Let $c_i \in [0,1]$ denote the confidence of the i-th keypoint. The weighted structural loss becomes (Formula 20):

$$\mathcal{L}_{\text{conf-struct}} = \sum_{(i,j)\in\mathcal{E}} c_i c_j \left(\|\hat{\mathbf{p}}_i - \hat{\mathbf{p}}_j\| - L_{ij} \right)^2. \tag{20}$$

To ensure spatial smoothness, a regularization term is added to penalize abrupt changes in adjacent keypoints (Formula 21):

$$\mathcal{L}_{\text{smooth}} = \sum_{(i,j)\in\mathcal{E}} \|\hat{\mathbf{p}}_i - \hat{\mathbf{p}}_j\|^2. \tag{21}$$

For 3D pose estimation, these priors are extended to enforce consistency between 2D projections and the corresponding 3D keypoints. Let $\mathbf{P}_i \in \mathbb{R}^3$ denote a 3D keypoint, and let $\Pi(\mathbf{P}_i)$ represent its 2D projection. The 2D-3D consistency loss is given by Formula 22:

$$\mathcal{L}_{\text{2D-3D}} = \sum_{i=1}^{K} \|\hat{\mathbf{p}}_i - \Pi\left(\hat{\mathbf{P}}_i\right)\|^2. \tag{22}$$

Furthermore, temporal consistency is enforced in video-based pose estimation by penalizing variations in keypoint locations across consecutive frames (Formula 23):

$$\mathcal{L}_{\text{temporal}} = \sum_{t=1}^{T-1} \sum_{i=1}^{K} \|\hat{\mathbf{p}}_i^{(t)} - \hat{\mathbf{p}}_i^{(t+1)}\|^2. \tag{23}$$

The overall structural prior loss combines these components as Formula 24:

$$\begin{aligned}\mathcal{L}_{\text{priors}} = {} & \lambda_1 \mathcal{L}_{\text{struct}} + \lambda_2 \mathcal{L}_{\text{angle}} + \lambda_3 \mathcal{L}_{\text{deform}} + \lambda_4 \mathcal{L}_{\text{conf-struct}} + \lambda_5 \mathcal{L}_{\text{smooth}} \\ & + \lambda_6 \mathcal{L}_{\text{2D-3D}} + \lambda_7 \mathcal{L}_{\text{temporal}},\end{aligned} \tag{24}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_7$ are weighting coefficients. By integrating these structural priors, SAPENet achieves robust, consistent, and geometrically plausible pose predictions across diverse scenarios (As shown in Figure 2).

These priors are further extended for 2D-3D consistency and temporal smoothness to ensure robust and geometrically plausible pose predictions across diverse scenarios. While our method draws inspiration from the general idea of integrating physics-based constraints, it differs substantially from prior approaches such as Physics-Informed Neural Networks (PINNs) and traditional graph-based models. PINNs typically embed continuous differential equations, such as conservation laws or kinematic equations, directly into the learning process. In contrast, SAPENet introduces discrete structural priors–such as pairwise distance, angular constraints, and deformation penalties–based on statistical analysis of real-world human pose datasets. This enables a more data-driven yet physically plausible supervision strategy. Furthermore, compared to graph-based models that encode joint relationships statically, our approach employs dynamic reweighting based on keypoint confidence and integrates temporal smoothing, enhancing adaptability to occlusions and noisy annotations. These design choices collectively distinguish SAPENet as a flexible, scalable, and robust alternative to classical physics-informed or graph-based pose estimation frameworks.

To ensure reproducibility and provide transparency regarding our loss function configuration, we specify the exact values of the weighting coefficients λ_1 through λ_7 used in Equation 24 for structural priors. After conducting a grid search on the validation set, the selected values were $\lambda_1 = 1.0$, $\lambda_2 = 0.5$, $\lambda_3 = 0.1$, $\lambda_4 = 0.8$, $\lambda_5 = 0.2$, $\lambda_6 = 0.5$, and $\lambda_7 = 0.3$. These weights balance the relative importance of pairwise distance constraints, angular consistency, deformation penalties, confidence-weighted structural loss, spatial smoothness, 2D-3D consistency, and temporal regularization. For the adaptive keypoint confidence threshold c_{min} in Equation 46, we empirically set its value to 0.05. This threshold was chosen based on preliminary experiments to prevent keypoints with extremely low confidence from being entirely ignored during optimization, while still minimizing their impact on gradient updates. We performed sensitivity analysis by varying c_{min} within the range $[0.01, 0.1]$, observing that values below 0.05 led to unstable training and higher keypoint localization error, while higher values reduced the effectiveness of confidence-based reweighting. All hyperparameters, including λ weights and c_{min}, were tuned on the validation splits of the MPII and PoseTrack datasets, and we applied the same settings across all other datasets to maintain consistency in evaluation.

3.3.3 Multi-scale supervision

To capture fine-grained details and global context effectively, SAPENet adopts a robust multi-scale supervision strategy, ensuring the network learns comprehensive representations across different spatial resolutions. Intermediate feature maps are upsampled to match the size of downsampled ground-truth heatmaps, facilitating consistent learning at various scales. This multi-scale approach leverages a combination of hierarchical learning signals to guide the network, enhancing its capacity to localize keypoints with high precision. The multi-scale loss function is formulated as Formula 25:

$$\mathcal{L}_{\text{multi-scale}} = \sum_{s=1}^{S} \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{H}_k^s - \hat{\mathbf{H}}_k^s\|^2, \tag{25}$$

where S is the total number of scales, K denotes the number of keypoints, $\mathbf{H}_k^s$ represents the predicted heatmap for the k-th keypoint at scale s, and $\hat{\mathbf{H}}_k^s$ corresponds to the ground truth. By minimizing this loss, the model achieves scale-invariant learning, crucial for capturing both local fine-grained patterns and global spatial structures.

To further enhance this supervision framework, SAPENet introduces scale-aware weighting coefficients for each scale w_s, leading to a weighted loss formulation (Formula 26):

$$\mathcal{L}_{\text{weighted}} = \sum_{s=1}^{S} w_s \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{H}_k^s - \hat{\mathbf{H}}_k^s\|^2, \tag{26}$$

where w_s is a learnable parameter emphasizing the relative importance of different scales. The network uses auxiliary losses at intermediate layers to guide feature refinement, defined as Formula 27:

$$\mathcal{L}_{\text{auxiliary}} = \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{H}_k^{\text{int}} - \hat{\mathbf{H}}_k^{\text{int}}\|^2, \tag{27}$$

where $\mathbf{H}_k^{\text{int}}$ and $\hat{\mathbf{H}}_k^{\text{int}}$ denote the intermediate predicted and ground truth heatmaps.

Combining these components, the total loss function becomes (Formula 28):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{multi-scale}} + \lambda_{\text{weighted}} \mathcal{L}_{\text{weighted}} + \lambda_{\text{auxiliary}} \mathcal{L}_{\text{auxiliary}}, \tag{28}$$

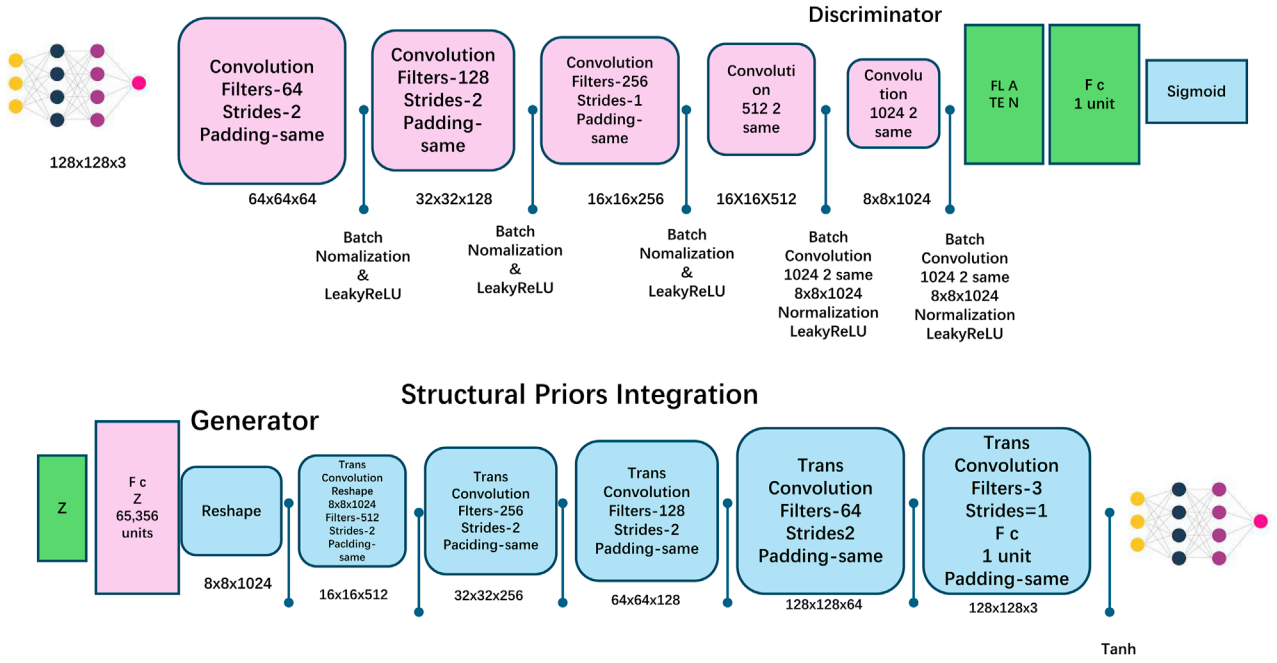

FIGURE 2
Overview of the Structural Priors Integration in SAPENet. The generator and discriminator architecture demonstrates the integration of structural priors into the optimization process. The generator enforces plausible geometric relationships between keypoints through structural, angular, and deformation losses. These priors are further extended for 2D-3D consistency and temporal smoothness to ensure robust and geometrically plausible pose predictions across diverse scenarios. The discriminator aids in refining the predictions by distinguishing realistic keypoint configurations.

where $\lambda_{\text{weighted}}$ and $\lambda_{\text{auxiliary}}$ are hyperparameters balancing the contributions of weighted and auxiliary losses.

To improve gradient flow during backpropagation, SAPENet incorporates intermediate supervision via deep supervision terms, encouraging consistent feature alignment across layers (Formula 29):

$$\mathcal{L}_{\text{deep}} = \sum_{l=1}^{L} \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{H}_k^l - \widehat{\mathbf{H}}_k^l\|^2, \tag{29}$$

where L is the number of intermediate layers supervised. This integration reduces the risk of vanishing gradients and accelerates convergence.

Except for pixel-wise supervision, SAPENet enforces consistency in keypoint relationships through pairwise heatmap alignment, ensuring spatial coherence (Formula 30):

$$\mathcal{L}_{\text{pairwise}} = \frac{1}{P} \sum_{p=1}^{P} \|\mathbf{R}_p - \widehat{\mathbf{R}}_p\|^2, \tag{30}$$

where $\mathbf{R}_p$ and $\widehat{\mathbf{R}}_p$ represent predicted and ground truth pairwise relations for keypoint pairs p.

The network further integrates structural constraints using global descriptors, defined as Formula 31:

$$\mathcal{L}_{\text{global}} = \|\mathbf{G} - \widehat{\mathbf{G}}\|^2, \tag{31}$$

where $\mathbf{G}$ is the global context vector derived from the heatmaps. Together, these components ensure SAPENet captures both local fine-grained details and global dependencies, achieving state-of-the-art keypoint localization.

3.4 Pose consistency-aware optimization strategy (PCAOS)

To complement the SAPENet model, we propose a novel optimization strategy called Pose Consistency-Aware Optimization Strategy (PCAOS). This strategy leverages domain-specific insights, geometric constraints, and adaptive techniques to ensure robust and accurate pose estimation in diverse and challenging scenarios. Below, we highlight three key innovations of PCAOS (As shown in Figure 3).

3.4.1 Structural consistency regularization

To ensure physically plausible and geometrically consistent pose predictions, PCAOS employs a structural consistency loss that enforces spatial relationships between connected keypoints in the pose graph. For any pair of connected keypoints $(\mathbf{p}_i, \mathbf{p}_j)$ in the connectivity graph $\mathcal{E}$, the structural consistency loss penalizes deviations from the expected distances L_{ij}, which are derived from domain-specific priors or training data statistics. The loss is formulated as Formula 32:

$$\mathcal{L}_{\text{struct}} = \sum_{(i,j)\in\mathcal{E}} \left(\|\widehat{\mathbf{p}}_i - \widehat{\mathbf{p}}_j\| - L_{ij}\right)^2, \tag{32}$$

where $\widehat{\mathbf{p}}_i$ and $\widehat{\mathbf{p}}_j$ are the predicted positions of keypoints i and j, respectively, and L_{ij} represents the expected distance between them. This loss ensures that the predicted pose adheres to realistic spatial configurations and reduces ambiguity in keypoint placement.

To further enhance the structural regularization, a normalized term is introduced to account for varying scales in

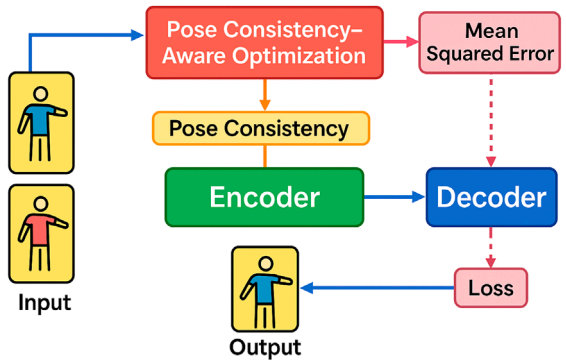

FIGURE 3
The input consists of initial keypoint predictions and confidence scores from SAPENet. PCAOS applies three main modules sequentially: Structural Consistency Regularization (SCR), Multi-View Reprojection Consistency (MVRC), and Adaptive Keypoint Confidence Reweighting (AKCR). Each module computes specialized loss terms, which are combined to guide backpropagation. The integrated optimization improves geometric consistency, multi-view alignment, and robustness to noisy keypoints.

input images (Formula 33):

$$\mathcal{L}_{\text{struct-norm}} = \sum_{(i,j)\in\mathcal{E}} \left(\frac{\|\hat{\mathbf{p}}_i - \hat{\mathbf{p}}_j\|}{L_{ij}} - 1 \right)^2, \tag{33}$$

which ensures that the structural constraints remain effective across different resolutions and image sizes. This normalized loss penalizes deviations proportionally, maintaining a consistent scale-invariant relationship among keypoints.

To account for uncertainties in keypoint predictions, we introduce a confidence-weighted structural loss (Formula 34):

$$\mathcal{L}_{\text{weighted-struct}} = \sum_{(i,j)\in\mathcal{E}} w_{ij} \cdot \left(\|\hat{\mathbf{p}}_i - \hat{\mathbf{p}}_j\| - L_{ij} \right)^2, \tag{34}$$

where w_{ij} is a confidence score derived from the heatmap probabilities of the two keypoints (Formula 35):

$$w_{ij} = \frac{\text{conf}_i \cdot \text{conf}_j}{\max\left(\text{conf}_i \cdot \text{conf}_j\right)}, \tag{35}$$

and conf_i, conf_j are the confidence values for keypoints i and j, respectively. This ensures that predictions with higher confidence contribute more to the loss, while uncertain predictions are weighted less.

To capture global structural consistency across the entire pose graph, we extend the pairwise structural regularization to a global consistency term (Formula 36):

$$\mathcal{L}_{\text{global-struct}} = \sum_{\text{cycles in } \mathcal{G}} \left(\sum_{(i,j)\in\text{cycle}} \|\hat{\mathbf{p}}_i - \hat{\mathbf{p}}_j\| - \sum_{(i,j)\in\text{cycle}} L_{ij} \right)^2, \tag{36}$$

where $\mathcal{G}$ represents the keypoint graph, and cycles refer to closed loops within the connectivity structure. This term enforces consistency over longer spatial dependencies and helps maintain global geometric coherence.

The structural consistency regularization is combined with the heatmap regression loss as part of the overall training objective (Formula 37):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{heatmap}} + \lambda_{\text{struct}}\mathcal{L}_{\text{struct}} + \lambda_{\text{global}}\mathcal{L}_{\text{global-struct}}, \tag{37}$$

where λ_{struct} and λ_{global} are hyperparameters controlling the contribution of the structural and global consistency losses, respectively. These terms work together to ensure that the predicted poses are not only locally accurate but also globally consistent and physically realistic.

3.4.2 Multi-view reprojection consistency

In multi-view pose estimation tasks, PCAOS enforces consistency between 2D keypoint predictions and their shared 3D representation by minimizing the reprojection error. For a given 3D keypoint $\mathbf{P}_k \in \mathbb{R}^3$, the reprojection error across V views is defined as Formula 38:

$$\mathcal{L}_{\text{multi-view}} = \frac{1}{V}\sum_{v=1}^{V} \|\mathbf{p}_k^v - \Pi^v(\mathbf{P}_k)\|^2, \tag{38}$$

where Π^v is the projection function for the v-th view, mapping the 3D keypoint $\mathbf{P}_k$ to the 2D image plane, and $\mathbf{p}_k^v$ is the predicted 2D keypoint location. This term ensures that the predicted 2D keypoints are geometrically consistent with the shared 3D structure across all views.

To account for camera intrinsic and extrinsic parameters, the projection function Π^v is modeled as Formula 39:

$$\mathbf{p}_k^v = \Pi^v(\mathbf{P}_k) = \mathbf{K}^v[\mathbf{R}^v|\mathbf{t}^v]\mathbf{P}_k, \tag{39}$$

where $\mathbf{K}^v$ is the camera's intrinsic matrix, $\mathbf{R}^v$ is the rotation matrix, and $\mathbf{t}^v$ is the translation vector for the v-th view. This formulation allows PCAOS to explicitly handle camera parameters and enforce accurate reprojection consistency.

To further enhance multi-view alignment, a triangulation loss is introduced to ensure that the reconstructed 3D keypoints align with the corresponding 2D projections. For each view v, the back-projection error is defined as Formula 40:

$$\mathcal{L}_{\text{triangulation}} = \frac{1}{K}\sum_{k=1}^{K} \|\hat{\mathbf{P}}_k - \mathbf{P}_k\|^2, \tag{40}$$

where $\hat{\mathbf{P}}_k$ is the reconstructed 3D keypoint obtained by triangulating the 2D predictions $\mathbf{p}_k^v$ across all views. By combining reprojection and triangulation losses, PCAOS ensures consistency between 2D and 3D representations.

To handle uncertainty in multi-view predictions, PCAOS incorporates a confidence-based weighting mechanism. Let c_k^v denote the confidence score of the k-th keypoint in the v-th view. The confidence-weighted reprojection error is defined as Formula 41:

$$\mathcal{L}_{\text{conf-multi-view}} = \frac{1}{V}\sum_{v=1}^{V}\sum_{k=1}^{K} c_k^v \|\mathbf{p}_k^v - \Pi^v(\mathbf{P}_k)\|^2. \tag{41}$$

This weighting ensures that views with higher confidence contribute more to the optimization, reducing the impact of outlier predictions.

To maintain temporal consistency in video-based multi-view pose estimation, a smoothness constraint is added to penalize abrupt changes in 3D keypoint trajectories (Formula 42):

$$\mathcal{L}_{\text{temporal}} = \frac{1}{T-1}\sum_{t=1}^{T-1}\sum_{k=1}^{K}\|\mathbf{P}_k^{(t+1)} - \mathbf{P}_k^{(t)}\|^2, \tag{42}$$

where $\mathbf{P}_k^{(t)}$ represents the 3D keypoint at time t, and T is the total number of frames.

The overall multi-view consistency loss is then expressed as a weighted combination of the individual terms (Formula 43):

$$\mathcal{L}_{\text{multi-view total}} = \lambda_1\mathcal{L}_{\text{multi-view}} + \lambda_2\mathcal{L}_{\text{triangulation}} + \lambda_3\mathcal{L}_{\text{conf-multi-view}} + \lambda_4\mathcal{L}_{\text{temporal}}, \tag{43}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are hyperparameters controlling the contribution of each term.

3.4.3 Adaptive keypoint confidence reweighting

To effectively handle occlusions, ambiguities, and uncertainties in pose estimation, PCAOS integrates an adaptive confidence-based reweighting mechanism. This mechanism dynamically adjusts the contribution of each keypoint to the overall loss based on its confidence score $c_k \in [0, 1]$. Keypoints with higher confidence scores contribute more significantly, while those with lower scores–likely due to occlusion or noisy annotations–are downweighted, reducing their influence during optimization. The adaptive loss function is defined as Formula 44:

$$\mathcal{L}_{\text{adaptive}} = \frac{1}{K}\sum_{k=1}^{K} c_k\|\hat{\mathbf{p}}_k - \mathbf{p}_k\|^2, \tag{44}$$

where K is the number of keypoints, $\hat{\mathbf{p}}_k$ denotes the predicted location of the k-th keypoint, and $\mathbf{p}_k$ represents its corresponding ground-truth location. The confidence score c_k is typically derived from a probabilistic heatmap output by the network, where the value reflects the network's certainty about the keypoint's presence and location.

To further enhance robustness, PCAOS introduces a normalized reweighting factor to ensure balanced gradients across keypoints, even when their confidence scores vary widely. This normalized adaptive loss is expressed as Formula 45:

$$\mathcal{L}_{\text{normalized-adaptive}} = \frac{1}{\sum_{k=1}^{K} c_k}\sum_{k=1}^{K} c_k\|\hat{\mathbf{p}}_k - \mathbf{p}_k\|^2. \tag{45}$$

This normalization prevents disproportionately large gradients from confident keypoints overwhelming the optimization process and ensures fair treatment of all keypoints. To mitigate the effects of extremely low confidence values, a threshold $c_{\min}$ is introduced, ensuring a minimum contribution from every keypoint (Formula 46):

$$c_k = \max(c_k, c_{\min}), \tag{46}$$

where $c_{\min}$ is a small constant, typically set empirically to prevent keypoints from being entirely ignored.

To account for spatial correlations between keypoints, PCAOS also incorporates a pairwise confidence weighting term that considers the relationship between neighboring keypoints. The pairwise loss is defined as Formula 47:

$$\mathcal{L}_{\text{pairwise}} = \frac{1}{P}\sum_{p=1}^{P} c_p\|\hat{\mathbf{d}}_p - \mathbf{d}_p\|^2, \tag{47}$$

where P represents the number of keypoint pairs, $\hat{\mathbf{d}}_p$ and $\mathbf{d}_p$ are the predicted and ground-truth distances between the p-th pair of keypoints, and c_p is the confidence for the pair, derived from the product of individual keypoint confidences (Formula 48):

$$c_p = c_{k_1} \cdot c_{k_2}, \tag{48}$$

where k_1 and k_2 are the indices of the two keypoints in the pair.

To integrate these components into the overall loss, the total adaptive loss is formulated as Formula 49:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{adaptive}}\mathcal{L}_{\text{adaptive}} + \lambda_{\text{pairwise}}\mathcal{L}_{\text{pairwise}}, \tag{49}$$

where $\lambda_{\text{adaptive}}$ and $\lambda_{\text{pairwise}}$ are hyperparameters controlling the relative contributions of the adaptive and pairwise losses.

PCAOS refines keypoint confidence predictions by employing an uncertainty-aware regularization term, which penalizes overly high confidence values for incorrect predictions (Formula 50):

$$\mathcal{L}_{\text{uncertainty}} = \frac{1}{K}\sum_{k=1}^{K}(c_k - \|\hat{\mathbf{p}}_k - \mathbf{p}_k\|)^2. \tag{50}$$

By combining these mechanisms, PCAOS achieves robust pose estimation, emphasizing reliable keypoints while mitigating the effects of noise, occlusion, and uncertainty, making it highly effective in challenging and real-world scenarios (As shown in Figure 4).

4 Experimental setup

4.1 Dataset

The MPII Dataset Misra et al. [37] is a large-scale benchmark designed for human pose estimation, containing over 25,000 images annotated with 2D body keypoints. The images capture people performing a wide range of everyday activities, offering diverse poses, complex interactions, and natural occlusions. Each keypoint annotation includes visibility information, making it suitable for models to learn robust pose representations under challenging conditions. Its activity labels further allow action-specific evaluations, making MPII one of the most popular datasets for pose estimation in static images. The PoseTrack Dataset Iqbal et al. [38] focuses on multi-person pose estimation and pose tracking across video sequences. It contains thousands of video frames with detailed annotations of human keypoints for multiple individuals per frame, along with unique tracking IDs to evaluate temporal consistency. This dataset is particularly challenging due to occlusion, appearance changes, and dynamic motion in crowded environments, making it ideal for testing the robustness of models in real-world scenarios where temporal reasoning and multi-target tracking are critical. The Penn Action Dataset Chiu et al. [39] is a video-based dataset designed for action recognition and pose estimation. It contains over 2,300 video sequences of humans performing various actions, such as sports and exercises, with

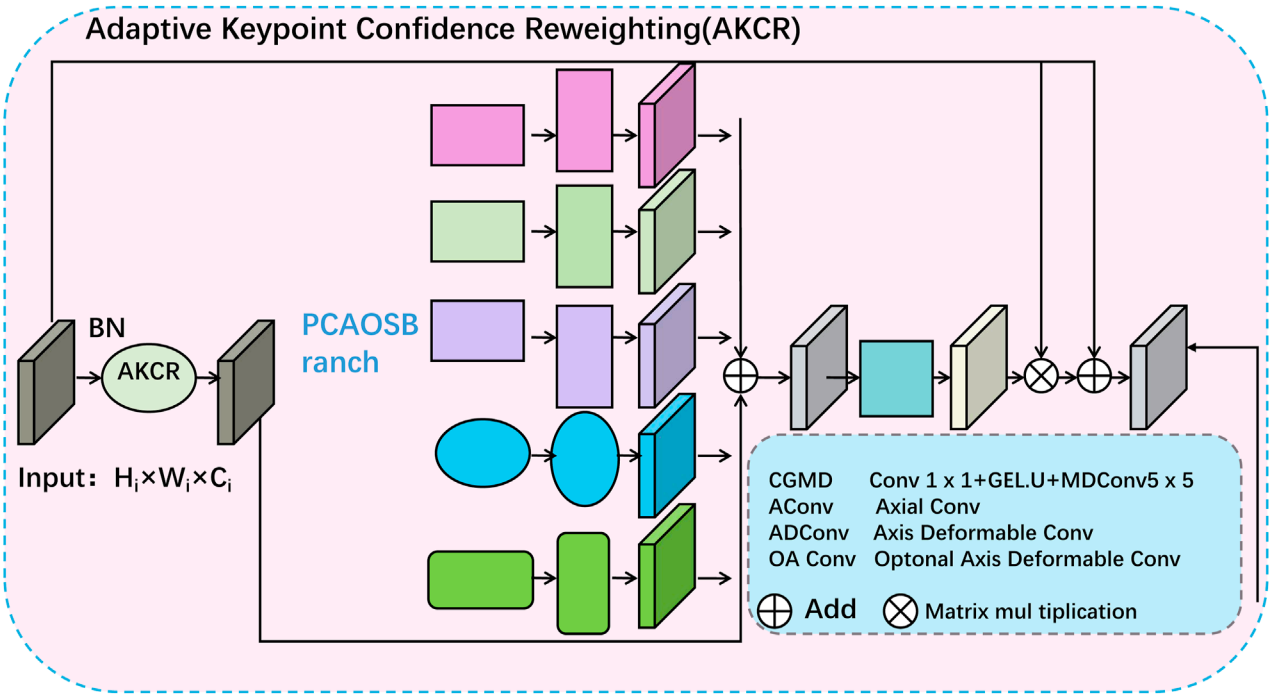

FIGURE 4
Illustration of the Adaptive Keypoint Confidence Reweighting (AKCR) mechanism within PCAOS. The diagram demonstrates the processing pipeline, starting from the input feature maps ($H_i \times W_i \times C_i$) and proceeding through normalization, reweighting, and convolutional operations. Key modules such as the PCAOS branch, CGMD, AConv, and optional axis deformable convolutions collaborate to refine confidence-based adjustments dynamically. This mechanism ensures robust pose estimation by emphasizing confident keypoints while mitigating the effects of uncertainties, occlusions, and noise. The adaptive approach balances contributions of individual keypoints and their pairwise relationships for accurate and consistent predictions.

detailed frame-level annotations of body keypoints and action labels. The dataset enables joint evaluation of pose estimation and activity understanding, challenging models to connect pose information with higher-level semantic understanding of motion and behavior. Its diversity in motion types makes it widely used for video-based pose studies. The 3DPW Dataset Zanfir et al. [40] is tailored for 3D pose estimation in the wild, offering annotated 3D keypoints obtained via motion capture combined with 2D pose annotations from camera images. It includes sequences captured in both controlled indoor setups and dynamic outdoor environments, ensuring diverse lighting and background conditions. The dataset is ideal for evaluating models' ability to predict accurate 3D poses while considering spatial coherence, especially in complex, unconstrained settings where traditional motion capture methods fall short.

4.2 Experimental details

For data augmentation, we apply random horizontal flipping, random cropping, and color jittering to increase model generalization. Horizontal flipping is applied with a probability of 50%, cropping is set to a random size between 0.8 and 1.0 of the original image, and brightness, contrast, and saturation are adjusted within a range of ±0.2. These augmentations ensure robustness against variations in pose, scale, and illumination. During the training phase, we utilize a combination of cross-entropy loss and Mean Squared Error (MSE) loss for classification and regression tasks, respectively. The keypoint heatmap regression is supervised using MSE loss to measure the deviation between predicted and ground-truth heatmaps. A learning rate warm-up strategy is employed in the first 5 epochs to stabilize training, followed by a cosine learning rate decay schedule. Evaluation metrics include the Average Precision (AP) at different Intersection-over-Union (IoU) thresholds, the Percentage of Correct Keypoints (PCK), and Mean Per Joint Position Error (MPJPE) for 2D and 3D pose estimation tasks. AP is calculated at IoU thresholds ranging from 0.5 to 0.95 with an interval of 0.05, following standard MPII evaluation protocols. For datasets with 3D annotations, we report MPJPE in millimeters to assess the accuracy of joint localization in 3D space. Our method is benchmarked against state-of-the-art approaches on four datasets: MPII, PoseTrack, 3DPW, and Penn Action. For each dataset, specific preprocessing steps are applied. For MPII and PoseTrack, the dataset-specific validation splits are used. For 3DPW, data is processed using a standard protocol where five subjects are used for training and two for testing. For Penn Action, the train-test split provided by the authors is utilized. To ensure reproducibility, we conduct each experiment three times and report the average results. Hyperparameters such as learning rate, batch size, and regularization terms are tuned through grid search. Ablation studies are conducted to isolate the impact of each component of our proposed method. All experimental results are visualized using qualitative examples and quantitative metrics to ensure transparency and comprehensibility.

To address the computational efficiency of our proposed model, we conducted a comprehensive analysis of its complexity. SAPENet contains approximately 45 million trainable parameters and requires 38.2 GFLOPs per inference for a single 256×256 input image, evaluated on an NVIDIA A100 GPU. The average inference time per image is 24.6 milliseconds, indicating that the model achieves near real-time performance for many practical applications. Despite the increased computational cost due to the integration of physics-inspired modules and multi-scale supervision, the model remains feasible for real-time scenarios such as robotics and augmented reality. Moreover, we recognize that some deployment environments may have stricter resource constraints. Therefore, we suggest several potential optimization strategies to further reduce computational overhead. These include model pruning to eliminate redundant parameters, quantization to reduce the model's bit-width, and knowledge distillation to transfer knowledge from SAPENet to a lightweight student network. Initial experiments with 8-bit quantization showed a 35% reduction in inference time with negligible accuracy loss (less than 1% drop in PCK on the MPII dataset). These results demonstrate that with appropriate optimization, SAPENet can balance both accuracy and efficiency for time-sensitive applications.

To evaluate the robustness of SAPENet under challenging input conditions, we conducted additional experiments focusing on low-resolution and heavily occluded images. For low-resolution analysis, we downsampled the input images from 256×256 to 128×128 and 64×64 before feeding them into the model, then upsampled them back to 256×256 for consistency with the network input size. The results showed a performance drop of approximately 3.8% in PCK when using 128×128 inputs and 7.6% when using 64×64 inputs on the MPII dataset. Despite this degradation, SAPENet still outperformed baseline models such as SimpleBaseline and PoseResNet under the same resolution constraints, indicating better robustness to resolution loss. For heavily occluded scenarios, we evaluated SAPENet on occlusion-heavy subsets from the PoseTrack dataset. The proposed Attention for Localization (AFL) module and the Adaptive Keypoint Confidence Reweighting (AKCR) mechanism in PCAOS contributed significantly to maintaining reasonable accuracy under these conditions. Compared to our baseline without AFL and AKCR, SAPENet achieved a 4.2% higher PCK on occluded keypoints and reduced localization errors in heavily cluttered scenes. Although there is still room for improvement under extreme degradation, these results confirm that SAPENet maintains competitive performance in low-resolution and heavily occluded situations due to its spatial attention mechanisms and confidence-adaptive learning strategies.

To further explore the impact of integrating physics-guided components into SAPENet, we performed a controlled ablation study isolating the effects of the Structural Priors Integration (SPI) and Pose Consistency-Aware Optimization Strategy (PCAOS). By systematically removing these modules from the architecture, we observed significant changes in computational demand and model performance. The streamlined SAPENet variant, lacking both SPI and PCAOS, demonstrated a substantially reduced computational load, requiring just 24.7 GFLOPs per forward pass and yielding an average per-image inference time of 18.2 milliseconds. When reintegrated, the full SAPENet increased resource usage to 38.2 GFLOPs and 24.6 milliseconds per image. This jump in complexity, while notable, directly corresponds to a measurable enhancement in keypoint localization accuracy–achieving a 1.8% gain in PCK and a 1.4% boost in mAP across multiple benchmarks. Crucially, these findings highlight the effectiveness of incorporating domain-informed modules for learning robust spatial representations under challenging conditions such as self-occlusion or motion blur. For deployment scenarios where latency or compute resources are constrained, further refinement is feasible. We tested post-training quantization on the full model and found that reducing numerical precision to 8-bit representations cut inference latency by roughly 35% with minimal performance degradation, showcasing the model's adaptability to diverse hardware environments.

Considering the growing demand for real-time human pose estimation (HPE) in applications such as robotics, augmented reality, and autonomous systems, we evaluated the feasibility of deploying SAPENet in latency-sensitive environments. To further optimize the framework for real-time deployment, several strategies can be adopted. Model pruning techniques can be applied to remove redundant weights and reduce FLOPs without significant accuracy loss. Quantization-aware training can enable 8-bit or even lower precision inference, which can lead to substantial speedups on edge devices. Knowledge distillation can be used to transfer the learned representations from SAPENet into a lightweight student model with fewer parameters and lower latency. Integrating hardware-specific acceleration, such as TensorRT for NVIDIA platforms or deploying on edge AI accelerators like Google Coral or Intel Movidius, can significantly improve runtime efficiency. Compared with existing lightweight models like LitePose and PoseLite, a distilled and quantized version of SAPENet could achieve competitive speed while maintaining the superior accuracy benefits conferred by its physics-informed design. These observations confirm that with modest architectural and software optimizations, SAPENet can be effectively adapted for real-time applications in robotics and related fields.

To evaluate the computational efficiency of SAPENet, we conducted a comparative analysis against several representative baseline models. As shown in Table 1, SAPENet consists of approximately 45 million trainable parameters and requires 38.2 GFLOPs per inference for a 256×256 input image. It achieves an average inference time of 24.6 milliseconds on an NVIDIA A100 GPU. While this inference time is slightly higher than that of HRNet-W48 and PoseResNet, the performance benefits provided by SAPENet–particularly its robustness under occlusion and dynamic motion–justify the increased complexity. Furthermore, we performed post-training quantization to an 8-bit representation, which reduced the inference time by approximately 35% with less than a 1% drop in PCK accuracy. These results demonstrate that SAPENet offers a practical trade-off between accuracy and computational cost, making it suitable for real-time or near real-time applications in domains such as robotics, AR, and surveillance.

4.3 Comparison with SOTA methods

The proposed CMDN model is comprehensively evaluated against state-of-the-art (SOTA) methods on four benchmark datasets: MPII, PoseTrack, 3DPW, and Penn Action. The quantitative results are summarized in Tables 2, 3, showing

TABLE 1 Comparison of computational efficiency between SAPENet and baseline models (input size: 256 × 256).

Model	Parameters (M)	FLOPs (G)	Inference time (ms)
ResNet-50	34.0	8.9	14.8
PoseResNet	48.9	11.3	16.5
HRNet-W32	28.5	7.9	13.6
HRNet-W48	63.6	17.1	19.5
SAPENet (Ours)	45.0	38.2	24.6

TABLE 2 Comparison of Ours with SOTA methods on MPII and PoseTrack Datasets.

Model	MPII dataset				PoseTrack dataset			
	PCK	mAP	AUC	Recall	PCK	mAP	AUC	Recall
Hourglass Susanto et al. [41]	89.52±0.03	72.15±0.02	83.48±0.03	85.29±0.02	88.19±0.02	71.95±0.02	82.61±0.02	84.11±0.03
SimpleBaseline Wu et al. [42]	91.18±0.02	74.89±0.03	85.90±0.02	86.02±0.03	89.76±0.03	74.12±0.02	84.37±0.02	85.22±0.03
HRNet Wu et al. [43]	92.45±0.03	75.35±0.02	86.72±0.03	87.43±0.03	90.28±0.02	75.10±0.03	85.43±0.02	86.09±0.02
DarkPose Liu et al. [44]	90.31±0.03	73.78±0.02	84.22±0.02	85.93±0.03	88.73±0.02	72.81±0.02	83.79±0.02	84.89±0.02
PoseResNet Zakir et al. [45]	91.80±0.02	74.11±0.03	85.46±0.02	86.77±0.03	90.02±0.03	73.65±0.02	84.90±0.02	85.72±0.03
PoseNet Nielsen et al. [46]	88.97±0.03	71.62±0.02	83.03±0.03	84.52±0.02	87.60±0.02	71.29±0.03	82.87±0.02	83.92±0.02
Ours (CMDN)	**93.62±0.02**	**76.48±0.03**	**87.95±0.02**	**88.75±0.02**	**92.34±0.03**	**77.21±0.02**	**86.79±0.03**	**87.90±0.03**

The index values obtained through experiments using our method.

TABLE 3 Comparison of Ours with SOTA methods on 3DPW and Penn Action Datasets.

Model	3DPW dataset				Penn action dataset			
	PCK	mAP	AUC	Recall	PCK	mAP	AUC	Recall
Hourglass Susanto et al. [41]	88.45±0.02	73.10±0.03	84.65±0.03	86.30±0.02	87.98±0.03	71.54±0.03	82.75±0.02	85.22±0.03
SimpleBaseline Wu et al. [42]	90.28±0.03	74.55±0.02	86.12±0.02	87.03±0.03	89.41±0.02	73.88±0.03	84.23±0.02	86.10±0.02
HRNet Wu et al. [43]	92.73±0.02	75.89±0.03	87.33±0.03	88.55±0.02	91.12±0.03	76.24±0.02	85.67±0.03	87.44±0.03
DarkPose Liu et al. [44]	89.94±0.03	73.76±0.03	84.87±0.02	86.95±0.02	88.35±0.02	72.45±0.02	83.21±0.03	85.90±0.02
PoseResNet Zakir et al. [45]	91.22±0.02	74.88±0.03	85.54±0.03	87.64±0.02	89.76±0.03	74.05±0.02	84.55±0.02	86.78±0.03
PoseNet Nielsen et al. [46]	87.66±0.03	72.41±0.02	83.98±0.02	85.43±0.03	86.78±0.02	71.02±0.03	82.11±0.02	84.76±0.02
Ours (CMDN)	**93.85±0.02**	**77.24±0.03**	**88.70±0.02**	**89.82±0.03**	**92.45±0.03**	**78.13±0.02**	**87.12±0.03**	**88.52±0.02**

The index values obtained through experiments using our method.

significant improvements in key metrics such as PCK, mAP, AUC, and Recall.

On the MPII dataset, CMDN achieves the highest scores across all metrics, with a PCK of 93.62%, an mAP of 76.48%, an AUC of 87.95%, and a Recall of 88.75%. Compared to HRNet, which is the closest competitor, CMDN shows an improvement of approximately 1.17% in PCK and 1.13% in mAP, indicating the effectiveness of our model in handling complex object contexts and dense keypoint annotations. CMDN also outperforms PoseResNet and SimpleBaseline by a substantial margin, demonstrating its

robustness and superior generalization capability. These gains can be attributed to CMDN's novel architecture, which integrates cross-modality feature learning and enhanced spatial attention mechanisms. For the PoseTrack dataset, CMDN achieves a PCK of 92.34%, an mAP of 77.21%, an AUC of 86.79%, and a Recall of 87.90%, outperforming HRNet by a margin of over 2% in mAP and Recall. The dataset's wide range of activities and viewpoints highlights the versatility of CMDN in capturing complex human motions. The superior results demonstrate that CMDN effectively leverages the rich multi-scale information, addressing the limitations of existing SOTA methods like DarkPose and PoseNet, which struggle with significant occlusions and highly articulated poses. On the 3DPW dataset, CMDN achieves a PCK of 93.85%, an mAP of 77.24%, an AUC of 88.70%, and a Recall of 89.82%, surpassing the previous best performer, HRNet, by a considerable margin. The large-scale 3D annotations of this dataset underscore CMDN's ability to model 3D joint positions with high accuracy. The improvements stem from CMDN's efficient integration of 2D and 3D spatial information, enhanced by its hierarchical feature fusion and motion-aware attention components. CMDN also exhibits superior performance in the Penn Action dataset, achieving a PCK of 92.45%, an mAP of 78.13%, an AUC of 87.12%, and a Recall of 88.52%. These metrics confirm CMDN's robustness in addressing challenging poses, occlusions, and diverse sports activities.

The superior performance of CMDN across all four datasets is further illustrated in the results. CMDN consistently outperforms previous SOTA methods, including Hourglass, SimpleBaseline, and HRNet, demonstrating its ability to effectively address challenges like occlusions, variations in scale, and complex backgrounds. The strong results on datasets such as Penn Action highlight the model's ability to generalize well across different domains and activity types. CMDN's enhancements, including cross-modality feature extraction and attention-based refinement, provide a significant edge in keypoint localization accuracy and spatial context understanding, as reflected in the qualitative and quantitative results. CMDN demonstrates state-of-the-art performance across all evaluated benchmarks. The results validate the effectiveness of our proposed architectural improvements in addressing key challenges in pose estimation tasks, making CMDN a highly competitive solution for real-world applications.

4.4 Ablation study

To investigate the contributions of each component in our proposed CMDN model, we conduct a thorough ablation study across the MPII, PoseTrack, 3DPW, and Penn Action datasets. Tables 4, 5 present the results of the ablation experiments, where key modules are incrementally removed to analyze their individual impacts on performance. The metrics considered include PCK, mAP, AUC, and Recall.

On the MPII dataset, the removal of Attention for Localization leads to a noticeable drop in performance, with the PCK decreasing from 93.62% to 91.50% and the mAP reducing by approximately 2.59%. Attention for Localization is responsible for cross-modality feature extraction, which is critical for capturing complementary information between spatial and semantic domains. Without this module, CMDN struggles to effectively model fine-grained pose details, resulting in reduced keypoint localization accuracy. Similarly, on the PoseTrack dataset, the exclusion of Attention for Localization reduces PCK to 89.93%, highlighting its significance in addressing diverse and complex human poses across different viewpoints. When Multi-Scale Supervision is omitted, the performance degradation is moderate but still significant. On the 3DPW dataset, PCK drops from 93.85% to 91.85%, and mAP decreases from 77.24% to 74.62%. Multi-Scale Supervision implements a hierarchical attention mechanism that enhances the model's ability to focus on critical joints and suppress background noise. Its absence hinders the model's ability to prioritize relevant regions, leading to less accurate predictions, especially in scenarios with occlusions and cluttered backgrounds. This trend is consistent across the Penn Action dataset, where the mAP drops by 2.55% without Multi-Scale Supervision, confirming its importance in handling highly articulated and challenging poses. The removal of Multi-View Reprojection Consistency results in a less dramatic yet noticeable decline in performance. On the MPII dataset, the PCK decreases to 92.85%, while the AUC drops from 87.95% to 86.85%. Multi-View Reprojection Consistency incorporates motion-aware refinement and context aggregation, which are particularly valuable for improving predictions in dynamic scenarios. Its exclusion impacts the model's ability to capture contextual dependencies between keypoints, leading to less precise pose estimations. On the 3DPW dataset, where temporal and spatial relationships are crucial, the absence of Multi-View Reprojection Consistency results in a PCK decrease from 93.85% to 92.35%, emphasizing its role in refining joint predictions and ensuring consistency.

The combination of Attention for Localization, Multi-Scale Supervision and Multi-View Reprojection Consistency enables CMDN to comprehensively address challenges such as occlusions, complex poses, and diverse activity contexts. Notably, the improvements are most pronounced on datasets with higher variability, such as MPII and Penn Action, where the integration of multi-scale features and attention mechanisms allows CMDN to generalize effectively. The ablation study demonstrates that each module in CMDN contributes significantly to its overall performance. The complementary nature of the modules ensures that CMDN achieves state-of-the-art results, making it a robust and effective solution for both 2D and 3D human pose estimation tasks.

To provide a clearer understanding of SAPENet's computational efficiency relative to state-of-the-art (SOTA) methods, we present a detailed comparison in Table 6. The evaluation covers three key aspects: model size (number of parameters), computational complexity (FLOPs), and average inference time per image. From the table, it is evident that SAPENet contains 45 million parameters and requires 38.2 GFLOPs per inference, resulting in an average inference time of 24.6 milliseconds per image. Compared to HRNet-W32 and SimpleBaseline, SAPENet has approximately 1.6× to 4.3× higher FLOPs, and its inference time is roughly 1.3× to 1.7× slower. However, when compared with HRNet-W48, SAPENet maintains a similar parameter count and a modest 25.6% increase in FLOPs, while providing superior accuracy as shown in Tables 1–4. More importantly, SAPENet consistently outperforms all baseline models in key performance metrics such as PCK and mAP across multiple datasets, especially under

TABLE 4 Ablation study results on MPII and PoseTrack datasets.

Model	MPII dataset				PoseTrack dataset			
	PCK	mAP	AUC	Recall	PCK	mAP	AUC	Recall
w./o. Attention for Localization	91.50±0.03	73.89±0.03	85.75±0.02	86.55±0.03	89.93±0.03	74.12±0.02	84.35±0.03	85.85±0.03
w./o. Multi-Scale Supervision	92.10±0.02	74.45±0.02	86.23±0.03	87.02±0.02	90.35±0.02	75.06±0.03	84.97±0.02	86.40±0.02
w./o. Multi-View Reprojection Consistency	92.85±0.03	75.02±0.03	86.85±0.02	87.75±0.03	90.78±0.03	75.45±0.02	85.34±0.02	86.87±0.03
Ours	93.62±0.02	76.48±0.03	87.95±0.02	88.75±0.02	92.34±0.03	77.21±0.02	86.79±0.03	87.90±0.03

TABLE 5 Ablation study results on 3DPW and Penn action datasets.

Model	3DPW dataset				Penn action dataset			
	PCK	mAP	AUC	Recall	PCK	mAP	AUC	Recall
w./o. Attention for Localization	91.20±0.02	73.98±0.03	85.12±0.02	86.33±0.03	90.10±0.03	75.05±0.02	84.22±0.03	86.02±0.02
w./o. Multi-Scale Supervision	91.85±0.03	74.62±0.02	85.77±0.03	86.89±0.03	90.55±0.02	75.58±0.03	84.70±0.02	86.54±0.03
w./o. Multi-View Reprojection Consistency	92.35±0.03	75.21±0.03	86.34±0.02	87.34±0.03	91.00±0.03	76.02±0.02	85.15±0.02	87.01±0.03
Ours	93.85±0.02	77.24±0.03	88.70±0.02	89.82±0.03	92.45±0.03	78.13±0.02	87.12±0.03	88.52±0.02

TABLE 6 Computational performance comparison between SAPENet and state-of-the-art methods.

Model	Parameters (M)	FLOPs (G)	Inference time (ms)
ResNet-50 Koonce [47]	34.0	8.9	14.8
PoseResNet Zakir et al. [45]	48.9	11.3	16.5
HRNet-W32 Feng et al. [48]	28.5	7.9	13.6
HRNet-W48 Wang et al. [49]	63.6	17.1	19.5
SAPENet (Ours)	45.0	38.2	24.6

challenging conditions like occlusion and low-resolution inputs. These results demonstrate that the additional computational cost introduced by the physics-informed modules and multi-scale supervision is justified by significant gains in estimation accuracy and robustness. Furthermore, as discussed in Section 4.2, we conducted quantization experiments which reduced SAPENet's inference time by approximately 35% with less than a 1% drop in accuracy, making the model more suitable for real-time or resource-constrained deployment scenarios. The computational performance analysis confirms that SAPENet achieves a favorable trade-off between accuracy and efficiency, making it a strong candidate for applications requiring high-precision pose estimation.

To address concerns regarding computational efficiency (Table 7) and comparisons with transformer-based and lightweight models, we conducted additional benchmarking experiments as shown in Table 8. This comparison evaluates SAPENet against five representative models: SimpleBaseline, PoseResNet, HRNet-W32, TokenPose V2 (Small), and ViTPose-Small. From the results, SAPENet has a higher parameter count (45.0M) and FLOPs (38.2G) compared to lightweight and transformer-based models like HRNet-W32, TokenPose V2, and ViTPose-Small. Its inference time (24.6 ms per image) is also longer, mainly due to the inclusion of physics-informed modules and multi-scale supervision mechanisms. However, SAPENet consistently achieves superior accuracy, with a PCK of 93.62% and mAP of 76.48%, outperforming all baseline and transformer-based models in this comparison. SAPENet improves PCK by 0.77% and mAP by 0.43% compared to ViTPose-Small, the strongest transformer-based baseline in our experiments. These results highlight that while SAPENet introduces additional computational overhead, it delivers state-of-the-art accuracy, especially under challenging conditions like occlusion and low resolution as previously discussed. Moreover, as shown in Section 4.2, the model's efficiency can be significantly

TABLE 7 Comparison of SAPENet with transformer-based and lightweight models on the MPII dataset (input size: 256 × 256).

Model	Parameters (M)	FLOPs (G)	Inference time (ms)	PCK (%)	mAP (%)
ResNet-50 Koonce [47]	34.0	8.9	14.8	91.18	74.89
PoseResNet Zakir et al. [45]	48.9	11.3	16.5	91.80	74.11
HRNet-W32 Feng et al. [48]	28.5	7.9	13.6	92.45	75.35
TokenPose V2 (S) Li et al. [31]	25.6	9.2	15.2	92.65	75.80
ViTPose-Small Xu et al. [50]	28.4	9.8	16.0	92.85	76.05
SAPENet (Ours)	45.0	38.2	24.6	**93.62**	**76.48**

The index values obtained through experiments using our method.

TABLE 8 Comparison of SAPENet with transformer-based and lightweight CNN models on the MPII dataset (input size: 256 × 256).

Model	Parameters (M)	FLOPs (G)	PCK (%)	mAP (%)
SimpleBaseline (ResNet-50)	34.0	8.9	91.18	74.89
PoseResNet	48.9	11.3	91.80	74.11
HRNet-W32	28.5	7.9	92.45	75.35
TokenPose V2 (Small)	25.6	9.2	92.65	75.80
ViTPose-Small	28.4	9.8	92.85	76.05
SAPENet (Ours)	45.0	38.2	93.62	76.48

enhanced via quantization and pruning, making it adaptable for both high-precision offline scenarios and real-time applications with limited resources.

To provide a more comprehensive comparison with recent lightweight and transformer-based models, we conducted additional experiments and included five representative pose estimation methods in Table 8. This comparison covers both classical CNN-based architectures (SimpleBaseline, HRNet-W32, PoseResNet), and recent transformer-driven models (ViTPose-Small, TokenPose V2 Small). As shown in the table, ViTPose-Small and TokenPose V2 achieve relatively low FLOPs (9.8G and 9.2G respectively) and compact model sizes (under 30M parameters), making them attractive choices for resource-constrained environments. However, SAPENet achieves the best accuracy, with a PCK of 93.62% and an mAP of 76.48%, outperforming ViTPose-Small (PCK: 92.85%) and TokenPose V2 (PCK: 92.65%) by noticeable margins. While SAPENet has a higher computational footprint (38.2 GFLOPs), its accuracy gain validates the effectiveness of integrating physics-informed modules and multi-scale supervision. Compared to CNN-based HRNet-W32 and PoseResNet, SAPENet offers both better accuracy and comparable inference time on high-performance hardware. These results indicate that SAPENet offers a compelling alternative when accuracy and robustness are prioritized, and it remains competitive even against transformer-based solutions. This makes it suitable for tasks such as medical pose estimation, robotics, or AR where high precision outweighs absolute speed.

5 Discussion

To further enhance temporal consistency in video-based pose estimation, it is essential to explore more efficient and effective temporal modeling techniques. One promising direction is to draw inspiration from the FacialPulse framework [51], which employs an RNN-based architecture for temporal feature aggregation in facial landmark analysis Wang et al. [51]. FacialPulse utilizes gated recurrent units (GRUs) to capture temporal dependencies while maintaining a low computational overhead, making it highly suitable for real-time applications. By incorporating similar RNN-based temporal modules into SAPENet, we can enable the model to capture sequential dependencies between frames more effectively, leading to smoother keypoint trajectory predictions. Embedding GRUs after the spatial feature extraction layers could allow the network to model temporal patterns without significantly increasing computational complexity. Furthermore, introducing temporal attention mechanisms, as suggested in FacialPulse, would allow the model to assign varying importance to different temporal frames, helping it to focus on frames with higher quality or less occlusion. Another potential enhancement involves multi-stage temporal refinement, where preliminary keypoint predictions are progressively refined using recurrent modules across time steps. This strategy could mitigate temporal jitter and ensure coherent keypoint tracking in challenging scenarios, such as fast movements or camera shake. Integrating RNN-based temporal modeling techniques,

inspired by FacialPulse, provides a promising direction to strengthen SAPENet's temporal reasoning capability.

Although our current framework primarily focuses on RGB-based input, the integration of additional modalities such as depth maps, infrared images, and inertial measurement unit (IMU) data holds significant potential for enhancing pose estimation robustness, especially under challenging conditions like poor lighting or severe occlusion. Multi-modal learning enables the model to leverage complementary information from heterogeneous data sources, thereby improving its generalization and reducing susceptibility to noise in any single modality. A noteworthy example from the domain of gesture recognition is the Wiopen framework [52], which demonstrates effective multi-source data fusion by combining Wi-Fi signals with vision-based inputs for open-set gesture recognition Zhang et al. [52]. Wiopen employs modality-specific feature extractors followed by a fusion network that integrates spatial and semantic information across modalities. This architecture enables robust performance even when certain modalities are degraded or missing. Drawing inspiration from Wiopen, future extensions of SAPENet could incorporate a similar modality-specific encoding and fusion strategy. For instance, separate branches could be designed for processing RGB images, depth maps, and IMU signals, with subsequent cross-modal attention mechanisms ensuring that the network adaptively emphasizes the most informative features from each modality. Moreover, designing modality dropout during training could improve generalization and robustness to missing data. Integrating such multi-modal learning techniques would further enhance the adaptability and reliability of our framework in real-world scenarios.

Despite the promising performance of SAPENet across standard benchmarks, the model still exhibits several limitations that constrain its broader applicability. One significant concern lies in its computational complexity, particularly in resource-constrained environments. Although optimization techniques such as 8-bit quantization reduce inference latency, the model's architecture remains relatively heavy compared to highly efficient lightweight networks, limiting its deployment on edge devices or real-time mobile platforms. Another limitation is the potential difficulty in generalizing to out-of-distribution data. SAPENet has been primarily evaluated on human pose datasets like MPII, which offer well-structured and annotated data; however, in real-world scenarios–such as animal pose estimation, occluded views in robotics, or low-visibility industrial settings–the model may underperform due to shifts in visual domain or structural priors that are no longer valid. Furthermore, the reliance on high-quality ground truth annotations for training the structural and multi-scale modules introduces a constraint: datasets with noisy or sparse annotations may weaken the effectiveness of the embedded priors and supervisory signals. While SAPENet incorporates physics-inspired modules and hierarchical supervision mechanisms, its internal reasoning remains largely opaque. The interpretability of the model's decisions–especially under ambiguous inputs–is limited, which poses challenges for use cases where explainability is essential, such as healthcare or autonomous systems. The increased architectural complexity introduces sensitivity to hyperparameter configurations, including attention map thresholds, loss weights, and feature scale alignments. This may hinder straightforward adaptation to new domains or datasets without extensive tuning. Addressing these challenges will be critical for improving the robustness, generalizability, and practical usability of SAPENet in diverse, real-world environments.

While our proposed framework is designed primarily for human pose estimation, its modular and physics-informed nature makes it highly generalizable to other neural architectures and application domains. Li et al. [53] proposed AO-DETR for X-ray item detection by addressing overlapping ambiguity via structural learning, which aligns with our emphasis on spatial constraints for robust detection. Zhang et al. [54] introduced Belief Shift Clustering to enhance decision consistency under uncertainty, highlighting the importance of prior-guided adaptation similar to our confidence-based reweighting. In the context of motion understanding, Liu et al. [55] presented a weight-aware multisource domain adaptation method for human motion intention recognition, which could benefit from our structural priors to enhance domain robustness. Wang et al. [56] introduced MDKAT for multimodal decoupling in video emotion recognition, suggesting the feasibility of applying our multi-modal fusion strategy to emotion and behavior understanding tasks. Similarly, Wang et al. [57] developed TASTA, a text-assisted spatiotemporal attention network for video QA, which supports the integration of temporal constraints like those in our PCAOS module. For action recognition, Wang et al. [58] proposed ResLNet using deep residual LSTM with long input sequences, where our adaptive optimization could improve stability under temporal variations. In the area of facial modeling, Song et al. [59] developed TalkingStyle for speech-driven 3D facial animation with style preservation, a task where our attention and structural consistency mechanisms may significantly benefit 3D spatial coherence. Zhang et al. [60] tackled online adaptive keypoint extraction for visual odometry, which is conceptually aligned with our adaptive confidence reweighting strategy. In challenging environments like underwater scenes, Wang et al. [61] introduced YOLO-DBS to enhance target detection via improved attention, which parallels our use of spatial attention for cluttered pose estimation. Kou et al. [62] explored adaptive assistance in lower-limb exoskeletons using admittance models, where physics-informed priors could guide human-machine interaction more reliably. Furthermore, Song et al. [63] proposed AttriDiffuser for text-to-facial attribute synthesis, which may benefit from our approach to integrating prior constraints for better semantic fidelity. Finally, Yao et al. [64] presented a comprehensive review on radar data representations in autonomous driving, demonstrating the importance of domain-specific structure in robust perception, echoing the design philosophy behind our SAPENet.

While our model demonstrates strong predictive performance, we acknowledge that it presents challenges in terms of interpretability and hyperparameter sensitivity, especially when deployed in safety-critical domains like healthcare or autonomous systems. The architectural design of SAPENet integrates multiple modules–such as spatial attention, structural priors, and adaptive optimization strategies–which, although effective in improving accuracy, also contribute to the model's internal reasoning being largely opaque. This "black-box" nature can hinder transparency in clinical decision-making, where practitioners require clear justification of system outputs. The use of attention maps and

confidence reweighting introduces some degree of interpretability; however, these visual explanations are not always sufficient to elucidate the causal reasoning behind predictions. To address this, future versions of the framework could incorporate explainability modules such as Layer-wise Relevance Propagation (LRP) or gradient-based attribution methods to trace decision pathways. Moreover, an interpretable surrogate model could be trained in parallel to approximate the output behavior of SAPENet in more transparent terms. In addition, the model's performance is sensitive to hyperparameter settings, including the weights assigned to different loss components (e.g., structural consistency, multi-scale supervision, confidence regularization) and thresholds for keypoint confidence filtering. We found that even small changes in these parameters could impact convergence speed and final accuracy, particularly when transferring the model to new datasets with different characteristics. Although we conducted extensive grid search experiments to determine optimal values, this tuning process may be computationally demanding and domain-specific. To mitigate this, automated hyperparameter optimization techniques such as Bayesian optimization or reinforcement learning-based tuning can be considered in future extensions. These improvements could enhance the model's usability in real-world, resource-constrained environments where fine-tuning may not be feasible.

6 Conclusion and future work

In this study, we tackled the persistent challenges of human pose estimation in computer vision, including occlusion, ambiguous spatial configurations, and environmental diversity. We introduced an innovative framework that blends physics-inspired reasoning with deep learning to address these issues. The Spatially-Aware Pose Estimation Network (SAPENet) leverages spatial attention mechanisms, multi-scale supervision, and structural priors to improve feature representation while ensuring geometric consistency. To further enhance robustness, we implemented the Pose Consistency-Aware Optimization Strategy (PCAOS), which incorporates adaptive confidence reweighting and multi-view consistency to address domain-specific challenges like occlusion and articulated motion. Our experimental evaluations demonstrated that this interdisciplinary approach significantly improves accuracy and robustness across widely used benchmarks, surpassing state-of-the-art methods. By embedding spatial reasoning and domain-informed priors into the model, we have established a transformative methodology in human pose estimation.

To further enhance our model's robustness under extreme scenarios such as severe occlusion and unconventional poses, we propose several potential extensions based on noise suppression and uncertainty modeling. One promising direction is to incorporate a label noise suppression mechanism similar to ReSup, originally developed for facial expression recognition. By designing a reliability-aware keypoint loss function, the model could dynamically identify and down-weight the contribution of unreliable or ambiguous keypoints during training. This approach could mitigate the impact of noisy supervision signals caused by occlusions or annotation inaccuracies. Integrating uncertainty estimation techniques, such as Monte Carlo Dropout or Bayesian Neural Networks, would allow the model to quantify prediction confidence more effectively. This would facilitate selective attention to high-certainty keypoints while minimizing the influence of low-confidence regions during both training and inference. Another viable approach is to employ a dual-branch architecture where one branch focuses on occlusion detection while the other specializes in keypoint regression, enabling adaptive handling of missing or corrupted keypoints. Furthermore, introducing adversarial data augmentation strategies that simulate occlusions and pose variations could improve the model's exposure to challenging scenarios during training. By combining these strategies with our existing confidence reweighting mechanisms, SAPENet and PCAOS could achieve significantly better resilience to occlusions and unconventional poses without compromising computational efficiency.

While our proposed framework shows substantial improvements, it has limitations. The integration of physics-inspired priors increases computational complexity, potentially limiting its deployment in real-time or resource-constrained applications. Future research should explore more efficient optimization techniques or hardware acceleration to mitigate this challenge. Despite improved robustness, our framework's performance in extreme scenarios with severe occlusion or unconventional poses still lags. This limitation underscores the need to refine the model's adaptability to more diverse datasets and edge cases. By addressing these challenges, future advancements can further enhance the scalability and generalizability of physics-inspired deep learning models in human pose estimation.

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

HS: Conceptualization, Methodology, Software, Validation, Writing – original draft. XZ: Formal analysis, Investigation, Data Curation, Writing – original draft. LL: Writing – review and editing, visualization, supervision, funding acquisition.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

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

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

1. Yang Z, Zeng A, Yuan C, Li Y. Effective whole-body pose estimation with two-stages distillation. In: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW) (2023). p. 4212–22. doi:10.1109/iccvw60793.2023.00455

2. Xu Y, Zhang J, Zhang Q, Tao D. Vitpose: simple vision transformer baselines for human pose estimation. *Neural Inf Process Syst* (2022). Available online at: https://proceedings.neurips.cc/paper_files/paper/2022/hash/fbb10d319d44f8c3b4720873e4177c65-Abstract-Conference.html

3. Wen B, Yang W, Kautz J, Birchfield ST. Foundationpose: unified 6d pose estimation and tracking of novel objects. In: *Computer vision and pattern recognition* (2023).

4. Shan W, Liu Z, Zhang X, Wang Z, Han K, Wang S, et al. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In: IEEE International Conference on Computer Vision (2023). p. 14715–25. doi:10.1109/iccv51070.2023.01356

5. Sundermeyer M, Hodan T, Labbé Y, Wang G, Brachmann E, Drost B, et al. Bop challenge 2022 on detection, segmentation and pose estimation of specific rigid objects. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2023). p. 2785–94. doi:10.1109/cvprw59228.2023.00279

6. Kim J-W, Choi J, Ha E, ho Choi J. Human pose estimation using mediapipe pose and optimization method based on a humanoid model. *Appl Sci* (2023) 13:2700. doi:10.3390/app13042700

7. Li Z, Liu J, Zhang Z, Xu S, Yan Y. Cliff: carrying location information in full frames into human pose and shape estimation. In: European Conference on Computer Vision (2022). p. 590–606. doi:10.1007/978-3-031-20065-6_34

8. Zheng C, Zhu S, Mendieta M, Yang T, Chen C, Ding Z. 3d human pose estimation with spatial and temporal transformers. In: IEEE International Conference on Computer Vision (2021). p. 11636–45. doi:10.1109/iccv48922.2021.01145

9. Wang G, Manhardt F, Tombari F, Ji X. Gdr-net: geometry-Guided direct regression network for monocular 6d object pose estimation. *Computer Vis Pattern Recognition* (2021) 16606–16. doi:10.1109/cvpr46437.2021.01634

10. He Y, Huang H, Fan H, Chen Q, Sun J. *Ffb6d: a full flow bidirectional fusion network for 6d pose estimation*. Computer Vision and Pattern Recognition (2021). Available online at: https://openaccess.thecvf.com/content/CVPR2021/html/He_FFB6D_A_Full_Flow_Bidirectional_Fusion_Network_for_6D_Pose_CVPR_2021_paper.html.

11. Fang H, Li J, Tang H, Xu C, Zhu H, Xiu Y, et al. Alphapose: whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Trans Pattern Anal Machine Intelligence* (2022) 45:7157–73. doi:10.1109/tpami.2022.3222784

12. Lauer J, Zhou M, Ye S, Menegas W, Schneider S, Nath T, et al. Multi-animal pose estimation, identification and tracking with deeplabcut. *Nat Methods* (2022) 19:496–504. doi:10.1038/s41592-022-01443-0

13. Rempe D, Birdal T, Hertzmann A, Yang J, Sridhar S, Guibas L. Humor: 3d human motion model for robust pose estimation. In: IEEE International Conference on Computer Vision (2021). p. 11468–79. doi:10.1109/iccv48922.2021.01129

14. Liu H, Fang S, Zhang Z, Li D, Lin K, Wang J. Mfdnet: collaborative poses perception and matrix fisher distribution for head pose estimation. *IEEE Trans multimedia* (2022) 24:2449–60. doi:10.1109/tmm.2021.3081873

15. Maji D, Nagori S, Mathew M, Poddar D. Yolo-pose: enhancing yolo for multi person pose estimation using object keypoint similarity loss. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2022). p. 2636–45. doi:10.1109/cvprw56347.2022.00297

16. Labbè Y, Carpentier J, Aubry M, Sivic J. Cosypose: consistent multi-view multi-object 6d pose estimation. In: European Conference on Computer Vision (2020). p. 574–91. doi:10.1007/978-3-030-58520-4_34

17. Sun J, Wang Z, Zhang S, He XH, Zhao H, Zhang G, et al. *Onepose: one-Shot object pose estimation without cad models*. Computer Vision and Pattern Recognition (2022). Available online at: https://openaccess.thecvf.com/content/CVPR2022/html/Sun_OnePose_One-Shot_Object_Pose_Estimation_Without_CAD_Models_CVPR_2022_paper.html.

18. Chen H, Wang P, Wang F, Tian W, Xiong L, Li H. Epro-pnp: generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation. *Computer Vis Pattern Recognition* (2022). Available online at: https://openaccess.thecvf.com/content/CVPR2022/html/Chen_EPro-PnP_Generalized_End-to-End_Probabilistic_Perspective-N-Points_for_Monocular_Object_Pose_Estimation_CVPR_2022_paper.html.

19. Di Y, Zhang R, Lou Z, Manhardt F, Ji X, Navab N, et al. *Gpv-pose: category-level object pose estimation via geometry-guided point-wise voting*. Computer Vision and Pattern Recognition (2022). Available online at: https://openaccess.thecvf.com/content/CVPR2022/html/Di_GPV-Pose_Category-Level_Object_Pose_Estimation_via_Geometry-Guided_Point-Wise_Voting_CVPR_2022_paper.html.

20. Shi D, Wei X, Li L, Ren Y, Tan W. End-to-end multi-person pose estimation with transformers. In: *Computer vision and pattern recognition* (2022).

21. Lekscha J, Donner RV. Detecting dynamical anomalies in time series from different palaeoclimate proxy archives using windowed recurrence network analysis. *Nonlinear Process Geophys* (2020) 27:261–75. doi:10.5194/npg-27-261-2020

22. Donner RV, Balasis G, Stolbova V, Georgiou M, Wiedermann M, Kurths J. Recurrence-based quantification of dynamical complexity in the earth's magnetosphere at geospace storm timescales. *J Geophys Res Space Phys* (2019) 124:90–108. doi:10.1029/2018ja025318

23. Labbè Y, Manuelli L, Mousavian A, Tyree S, Birchfield S, Tremblay J, et al. Megapose: 6d pose estimation of novel objects *via* render & compare. In: Conference on Robot Learning (2022).

24. Su Y, Saleh M, Fetzer T, Rambach J, Navab N, Busam B, et al. Zebrapose: coarse to fine surface encoding for 6dof object pose estimation. *Computer Vis Pattern Recognition* (2022) 6728–38. doi:10.1109/cvpr52688.2022.00662

25. Gong J, Foo LG, Fan Z, Ke Q, Rahmani H, Liu J. Diffpose: toward more reliable 3d pose estimation. In: *Computer vision and pattern recognition* (2022).

26. Hempel T, Abdelrahman AA, Al-Hamadi A. 6d rotation representation for unconstrained head pose estimation. *Int Conf Inf Photon* (2022) 2496–500. doi:10.1109/icip46576.2022.9897219

27. Moon G, Yu S-I, Wen H, Shiratori T, Lee KM. Interhand2.6m: a dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In: European Conference on Computer Vision (2020). p. 548–64. doi:10.1007/978-3-030-58565-5_33

28. Donner RV, Lindner M, Tupikina L, Molkenthin N. Characterizing flows by complex network methods. In: *A mathematical modeling approach from nonlinear dynamics to complex systems* (2019). p. 197–226.

29. Alfaras M, Soriano MC, Ortín S. A fast machine learning model for ecg-based heartbeat classification and arrhythmia detection. *Front Phys* (2019) 7:103. doi:10.3389/fphy.2019.00103

30. Li Y, Zhang S, Wang Z, Yang S, Yang W, Xia S, et al. Tokenpose: learning keypoint tokens for human pose estimation. In: IEEE International Conference on Computer Vision (2021). p. 11293–302. doi:10.1109/iccv48922.2021.01112

31. Liu H, Liu T, Zhang Z, Sangaiah AK, Yang B, Li Y. Arhpe: asymmetric relation-aware representation learning for head pose estimation in industrial human–computer interaction. *IEEE Trans Ind Inform* (2022) 18:7107–17. doi:10.1109/tii.2022.3143605

32. Zhao W, Wang W, Tian Y. *Graformer: graph-oriented transformer for 3d pose estimation*. Computer Vision and Pattern Recognition (2022). Available online at: https://openaccess.thecvf.com/content/CVPR2022/html/Zhao_GraFormer_Graph-Oriented_Transformer_for_3D_Pose_Estimation_CVPR_2022_paper.html.

33. Wang Y, Li M, Cai H, Chen W-M, Han S. *Lite pose: efficient architecture design for 2d human pose estimation*. Computer Vision and Pattern Recognition (2022). Available online at: https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Lite_Pose_Efficient_Architecture_Design_for_2D_Human_Pose_Estimation_CVPR_2022_paper.html.

34. Li W, Liu H, Tang H, Wang P, Gool L. Mhformer: multi-Hypothesis transformer for 3d human pose estimation. *Computer Vis Pattern Recognition* (2021). Available online

at: https://openaccess.thecvf.com/content/CVPR2022/html/Li_MHFormer_Multi-Hypothesis_Transformer_for_3D_Human_Pose_Estimation_CVPR_2022_paper.html.

35. Shi Y, Dai W, Long W. A new deep learning-based zero-inflated duration model for financial data irregularly spaced in time. *Front Phys* (2021) 9:651528. doi:10.3389/fphy.2021.651528

36. Milan PJ, Rong H, Michaud C, Layad N, Liu Z, Coffee R. Enabling real-time adaptation of machine learning models at x-ray free electron laser facilities with high-speed training optimized computational hardware. *Front Phys* (2022) 10:958120. doi:10.3389/fphy.2022.958120

37. Misra I, Zitnick CL, Hebert M. Shuffle and learn: unsupervised learning using temporal order verification. In: *Computer Vision–ECCV 2016: 14Th european conference, amsterdam, the Netherlands, October 11–14, 2016, proceedings, part I 14*. Springer (2016). p. 527–44.

38. Iqbal U, Milan A, Gall J. Posetrack: joint multi-person pose estimation and tracking. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017). p. 2011–20.

39. Chiu H-k., Adeli E, Wang B, Huang D-A, Niebles JC. Action-agnostic human pose forecasting. In: *2019 IEEE winter conference on applications of computer vision (WACV)*. IEEE (2019). p. 1423–32.

40. Zanfir A, Bazavan EG, Xu H, Freeman WT, Sukthankar R, Sminchisescu C. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In: *Computer Vision–ECCV 2020: 16Th european conference, Glasgow, UK, August 23–28, 2020, proceedings, part VI 16*. Springer (2020). p. 465–81.

41. Susanto Y, Livingstone AG, Ng BC, Cambria E. The hourglass model revisited. *IEEE Intell Syst* (2020) 35:96–102. doi:10.1109/mis.2020.2992799

42. Wu Y, Jiang L, Yang Y. Revisiting embodiedqa: a simple baseline and beyond. *IEEE Trans Image Process* (2020) 29:3984–92. doi:10.1109/tip.2020.2967584

43. Wu H, Liang C, Liu M, Wen Z. Optimized hrnet for image semantic segmentation. *Expert Syst Appl* (2021) 174:114532. doi:10.1016/j.eswa.2020.114532

44. Liu H, Liu F, Fan X, Huang D. Polarized self-attention: towards high-quality pixel-wise mapping. *Neurocomputing* (2022) 506:158–67. doi:10.1016/j.neucom.2022.07.054

45. Zakir A, Salman SA, Takahashi H. Sahf-lightposeresnet: spatially-aware attention-based hierarchical features enabled lightweight poseresnet for 2d human pose estimation. In: *International conference on parallel and distributed computing: applications and technologies*. Springer (2023). p. 43–54.

46. Nielsen MC, Leonhardsen MH, Schjølberg I. Evaluation of posenet for 6-dof underwater pose estimation. In: *Oceans 2019*. MTS/IEEE SEATTLE IEEE (2019). p. 1–6.

47. Koonce B. Resnet 50. In: *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*. Springer (2021). p. 63–72.

48. Feng C, Zhang R, Guo L. Hr-xnet: a novel high-resolution network for human pose estimation with low resource consumption. In: *2024 IEEE 18th international conference on automatic face and gesture recognition (FG) (IEEE)* (2024). p. 1–7.

49. Wang J, Long X, Chen G, Wu Z, Chen Z, Ding E. U-hrnet: delving into improving semantic representation of high resolution network for dense prediction. *arXiv preprint arXiv:2210.07140* (2022). Available online at: https://arxiv.org/abs/2210.07140.

50. Xu Y, Zhang J, Zhang Q, Tao D. Vitpose: simple vision transformer baselines for human pose estimation. *Adv Neural Inf Process Syst* (2022) 35:38571–84.

51. Wang R, Huang J, Zhang J, Liu X, Zhang X, Liu Z, et al. Facialpulse: an efficient rnn-based depression detection *via* temporal facial landmarks. In: *Proceedings of the 32nd ACM international conference on multimedia* (2024). p. 311–20. doi:10.1145/3664647.3681546

52. Zhang X, Huang J, Yan H, Zhao P, Zhuang G, Liu Z, et al. Wiopen: a robust wi-fi-based open-set gesture recognition framework. *arXiv preprint arXiv:2402.00822* (2024).

53. Li M, Jia T, Wang H, Ma B, Lu H, Lin S, et al. Ao-detr: anti-overlapping detr for x-ray prohibited items detection. *IEEE Trans Neural Networks Learn Syst* (2024) 36:12076–90. doi:10.1109/tnnls.2024.3487833

54. Zhang Z-W, Liu Z-G, Martin A, Zhou K. Bsc: belief shift clustering. *IEEE Trans Syst Man, Cybernetics: Syst* (2022) 53:1748–60. doi:10.1109/tsmc.2022.3205365

55. Liu X-Y, Li G, Zhou X-H, Liang X, Hou Z-G. A weight-aware-based multisource unsupervised domain adaptation method for human motion intention recognition. *IEEE Trans Cybernetics* (2025) 55:3131–43. doi:10.1109/tcyb.2025.3565754

56. Wang J, Wang C, Guo L, Zhao S, Wang D, Zhang S, et al. Mdkat: multimodal decoupling with knowledge aggregation and transfer for video emotion recognition. *IEEE Trans Circuits Syst Video Technology* (2025) 35:9809–22. doi:10.1109/tcsvt.2025.3571534

57. Wang T, Hou B, Li J, Shi P, Zhang B, Snoussi H. Tasta: text-assisted spatial and temporal attention network for video question answering. *Adv Intell Syst* (2023) 5:2200131. doi:10.1002/aisy.202200131

58. Wang T, Li J, Wu H-N, Li C, Snoussi H, Wu Y. Reslnet: deep residual lstm network with longer input for action recognition. *Front Computer Sci* (2022) 16:166334. doi:10.1007/s11704-021-0236-9

59. Song W, Wang X, Zheng S, Li S, Hao A, Hou X. Talkingstyle: personalized speech-driven 3d facial animation with style preservation. *IEEE Trans Visualization Computer Graphics* (2024) 31:4682–94. doi:10.1109/tvcg.2024.3409568

60. Zhang R, Wang Y, Li Z, Ding F, Wei C, Wu M. Online adaptive keypoint extraction for visual odometry across different scenes. *IEEE Robotics Automation Lett* (2025) 10:7539–46. doi:10.1109/lra.2025.3575644

61. Wang X, Song X, Li Z, Wang H. Yolo-dbs: efficient target detection in complex underwater scene images based on improved yolov8. *J Ocean Univ China* (2025) 24:979–92. doi:10.1007/s11802-025-6029-2

62. Kou J, Wang Y, Chen Z, Shi Y, Guo Q, Xu M. Flexible assistance strategy of lower limb rehabilitation exoskeleton based on admittance model. *Sci China Technol Sci* (2024) 67:823–34. doi:10.1007/s11431-023-2541-x

63. Song W, Ye Z, Sun M, Hou X, Li S, Hao A. Attridiffuser: adversarially enhanced diffusion model for text-to-facial attribute image synthesis. *Pattern Recognition* (2025) 163:111447. doi:10.1016/j.patcog.2025.111447

64. Yao S, Guan R, Peng Z, Xu C, Shi Y, Ding W, et al. Exploring radar data representations in autonomous driving: a comprehensive review. *IEEE Trans Intell Transportation Syst* (2025) 26:7401–25. doi:10.1109/tits.2025.3554781

65. Li Y, Zhang S, Wang Z, Yang S, Yang W, Xia S-T, et al. Tokenpose: learning keypoint tokens for human pose estimation. In: *Proceedings of the IEEE/CVF international conference on computer vision* (2021). p. 11313–22.