frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 29 October 2025
DOI 10.3389/fmars.2025.1678417

OPEN ACCESS

EDITED BY
Yakun Ju,
University of Leicester, United Kingdom

REVIEWED BY
Navid Nourani-Vatani,
Imperium Drive Ltd, United Kingdom
Yize Wang,
Waseda University, Japan

*CORRESPONDENCE
Xing Peng
pengxing22@nudt.edu.cn

RECEIVED 02 August 2025
ACCEPTED 29 September 2025
PUBLISHED 29 October 2025

CITATION
Li S and Peng X (2025) DyAqua-YOLO: a high-precision real-time underwater object detection model based on dynamic adaptive architecture.
Front. Mar. Sci. 12:1678417.
doi: 10.3389/fmars.2025.1678417

DyAqua-YOLO: a high-precision real-time underwater object detection model based on dynamic adaptive architecture

Shucheng Li[1,2] and Xing Peng[1,2,3]*

[1]National Key Laboratory of Equipment State Sensing and Smart Support, Changsha, Hunan, China, [2]College of Intelligent Science and Technology, National University of Defense Technology, Changsha, Hunan, China, [3]Hunan Provincial Key Laboratory of Ultra-Precision Machining Technology, Changsha, Hunan, China

With the rapid development of marine resource exploitation and the increasing demand for underwater robot inspection, achieving reliable target perception in turbid, low-illumination, and spectrally limited underwater environments has become a key challenge that urgently needs to be addressed in the field of computer vision. This paper proposes DyAqua-YOLO, a dynamic adaptive model specifically designed to address the critical challenges of low-contrast blurred targets and pervasive small-object detection in complex underwater optical environments. Central to our approach are three core innovations: 1) The Dynamic Scale Sequence Feature Fusion (DySSFF) module, which replaces static upsampling with a dynamic grid generator to preserve spatial details of blurred and small targets; 2) The DyC3k2 module, which introduces dynamic kernel weighting into the reparameterization process, enabling adaptive feature extraction for degraded underwater images; 3) A unified Focaler-Wise Normalized Wasserstein Distance (FWNWD) loss, not a mere combination but a hierarchical framework where WIoU provides gradient modulation, Focaler-IoU handles hard-easy sample bias, and NWD ensures small-object sensitivity, working in concert to resolve optimization conflicts. On the DUO dataset containing 74,515 instances, the DyAqua-YOLO model achieves mAP@0.5 of 91.8% and mAP@[0.5:0.95] of 72.2%, demonstrating outstanding accuracy. Compared to the baseline (YOLO11n), these metrics have improved by 3.9% and 3.7%, respectively. On the OrangePi AIpro platform (8TOPS NPU, 16GB RAM), the enhanced model achieves an inference speed of 21 FPS, striking an optimal balance between accuracy and efficiency. Ablation experiments show that the DyC3k2 module increases mAP@0.5 by 1.2% and mAP@[0.5:0.95] by 1.7% compared to the YOLO11 baseline model, while reducing FLOPs by 3.2%, thereby enhancing model accuracy and optimizing computational efficiency. The FWNWD loss function improves the recall of small targets by 3.6% compared to the CIoU loss function, effectively balancing the optimization conflicts between hard examples and small targets and improving localization accuracy. This research provides a new approach for high-precision real-time detection in

underwater embedded devices, and its dynamic and adjustable architecture has broad applicability guiding value for application in other scenarios with similar challenges.

KEYWORDS

underwater object detection, deep learning, YOLO, FWNWD, dynamic adaptive model, high-precision

1 Introduction

Underwater object detection, as a core technology in fields such as marine resource exploration, underwater robot navigation, and ecological monitoring, has received extensive attention in recent years. With the increasing global demand for marine resource development [according to the United Nations report on ocean affairs, the scale of the ocean economy is expected to reach 3 trillion US dollars by 2030 (United Nations, 2022)], efficient and accurate underwater target recognition algorithms have become a key bottleneck restricting the application of this technology. However, the complex optical environment underwater leads to significant degradation in image quality. Irregular light scattering and differential absorption in water cause severe color distortion and contrast attenuation in underwater images (Liu et al., 2020). The low detectability of underwater targets poses a dual challenge of high false detection and missed detection rates for underwater object detection. At the same time, to ensure the mobility and flexibility of underwater detectors, the computing power of underwater embedded devices is greatly constrained, making it difficult to meet the requirements of rapidity and real-time performance for underwater object detection. In summary, designing an accurate, fast, and lightweight underwater object detection model has become a challenging issue.

Currently, object detection technologies include traditional target recognition algorithms and deep learning-based object detection algorithms (Xu et al., 2023). Traditional target recognition algorithms use feature extractors to extract image features, such as Histogram of Oriented Gradients (HOG) (Dalal and Triggs, 2005) and Viola-Jones detector (Viola and Jones, 2001). Compared with traditional target recognition algorithms, deep learning-based object detection algorithms, starting with R-CNN (Girshick et al., 2014), have established multi-structured network models and designed adaptive feature extraction algorithms, effectively improving detection speed, accuracy, and robustness in complex environments (Amjoud and Amrouch, 2023). Since then, the field of object detection has continued to make new breakthroughs, with the emergence of algorithms such as Fast R-CNN (Girshick, 2015), Faster R-CNN (Ren et al., 2017), SPPNet (He et al., 2015), R-FCN (Dai et al., 2016), SSD (Liu et al., 2016), and YOLO (You Only Look Once; Redmon et al., 2016) series (Zaidi et al., 2022). YOLO (Redmon et al., 2016), as a single-stage object detector, achieves efficient real-time detection by directly regressing bounding box coordinates and class probabilities, demonstrating great potential in the embedded deployment of underwater equipment. However, due to the complex underwater environment, such as light scattering, low contrast, and noise interference, traditional YOLO models face numerous challenges in underwater object detection tasks (Luo and Feng, 2025; Zheng and Yu, 2024). Based on this, researchers have proposed a series of improvement strategies.

Feature extraction is one of the core steps in object detection. In underwater environments, the complex background and low-quality images make it difficult for traditional feature extraction methods to effectively capture the key information of the target (Luo and Feng, 2025; Zheng and Yu, 2024). Therefore, many studies have focused on optimizing the design of the feature extraction module to enhance the model's ability to recognize underwater targets. A common approach is to introduce attention mechanisms, such as CBAM (Convolutional Block Attention Module), which can adaptively adjust the weights in the channel and spatial dimensions, thereby highlighting the information in important regions (Luo and Feng, 2025; Zheng and Yu, 2024). Another approach is to utilize the dynamic sparse attention mechanism (BiFormer) combined with the Ghost Bottleneck module, further reducing computational costs and improving detection accuracy (Chen et al., 2024; Zhang et al., 2024). To address the wide distribution of target sizes in underwater scenarios, multi-scale feature fusion technology has become a research focus (Lu et al., 2024). RG-YOLO integrates the GDFPN feature pyramid network, the reparameterized multi-scale fusion module (RMF), and the dynamic head module, effectively aggregating cross-level features, thereby enhancing the model's detection ability for small and dense targets, achieving an mAP@0.5 of 86.1% on the DUO dataset (Zheng and Yu, 2024). Similarly, MarineYOLO improves the Feature Pyramid Network (FPN) and incorporates the Efficient Multi-scale Attention (EMA) module in the backbone network, enhancing the fine-grained expression of small targets through global modeling of feature channels (Liu et al., 2024), achieving an average accuracy of 88.1% on the URPC dataset and 78.5% on the RUOD dataset. Additionally, some works utilize spatial pyramid pooling (SPP) or lightweight convolutional modules to improve the utilization of multi-scale features (Cheng et al., 2024; Luo et al., 2024). By introducing the SPP module or multi-layer perceptron (MLP), the model's adaptability to targets of

different scales can be effectively enhanced (Liu et al., 2024). Meanwhile, some studies have attempted to combine traditional attention mechanisms with deformable convolutions to better capture information in key regions (Luo and Feng, 2025; Zheng and Yu, 2024). To address the problem of high proportion of small targets and easy missed detections in underwater scenes, scholars generally take the following measures: introducing shape-sensitive similarity metrics (such as Shape-IoU), and designing specialized modules for capturing fine-grained features (such as OMNI-Dynamic Convolution) (Lu et al., 2024; Cheng et al., 2024). Moreover, the dynamic sparse attention mechanism not only reduces computational costs but also significantly improves the detection accuracy of small targets (Zheng et al., 2024). In the actual deployment process, especially in embedded devices, the limited computing resources make traditional high-parameter models difficult to apply. Therefore, developing algorithms that can maintain high detection accuracy while having low computational costs becomes particularly important. YOLO_GN builds a lightweight backbone network based on GhostNetV2 and combines the sparse attention mechanism BiFormer, significantly reducing the computational cost (Chen et al., 2024). It demonstrates strong application potential on embedded devices, achieving a detection accuracy of 85.35% when training the URPC dataset on the Raspberry Pi 4B platform, far exceeding the performance of similar products (Chen et al., 2024; Zhang et al., 2024). Similarly, RTL-YOLOv8n, through means such as the lightweight coupled detection head (LCD-Head), reduces the computational load by 31.6% compared to the original YOLOv8n model while increasing mAP@0.5 by 1.5%, successfully achieving a good balance between performance and efficiency (Feng et al., 2024). Additionally, studies have shown that replacing standard convolution operations with grouped convolution or depthwise convolution can effectively reduce the number of parameters and accelerate the inference process (Zhang et al., 2024; Hu et al., 2025).

This paper proposes an underwater object detection model based on YOLO11n. The main contributions of this paper are as follows:

1. To address the issue of multi-scale features of underwater targets, the ASF-YOLO framework is applied to the YOLO11 model. By combining spatial and scale features, the detection and performance in scenarios with small and dense targets are significantly enhanced.
2. In the Scale Sequence Feature Fusion (SSFF) module of ASF-YOLO, DySample lightweight dynamic upsampling is used to replace the traditional linear interpolation upsampling, and Dynamic Scale Sequence Feature Fusion (DySSFF) is proposed. This improves the multi-scale feature fusion ability and the detection performance of small targets without significantly increasing the computational burden.
3. The DyC3k2 feature extraction module is designed to enhance the model's adaptability to complex underwater environments through dynamic convolution kernel weight allocation, thereby improving the model's feature extraction ability.
4. In the training stage, a joint loss optimization strategy is adopted, which combines the WIoU dynamic focusing mechanism and Focaler-IoU bounding box loss, and integrates the NWD metric for small targets. A novel FWNWD loss function is proposed, which improves the accuracy of the model when training on underwater datasets and the recall of small object detection.

Experiments show that on the DUO dataset, DyAqua-YOLO achieves 91.8% mAP@0.5 and 72.2% mAP@[0.5:0.95], with an inference speed of 21 FPS (OrangePi AIpro, 8TOPS NPU, 16GB RAM), meeting the real-time requirements. To systematically explain this work, the subsequent structure of the paper is as follows: The Methodology section elaborates on the overall architecture of the DyAqua-YOLO models and the design principles of its core innovative modules - Dynamic Scale Sequence Feature Fusion (DySSFF), dynamic convolution (DyC3k2), and the joint loss function (FWNWD); the Experiment and Discussion section reports detailed experimental settings, result comparisons, and ablation analyses based on the DUO dataset, quantitatively verifying the effectiveness and superiority of the model; the Conclusion section summarizes the core findings and contributions, and discusses future directions.

2 Methodology

2.1 Overall framework design

YOLO11 (Khanam and Hussain, 2024, preprint) is a new generation of object detection algorithm developed by Ultralytics based on the YOLOv8 architecture. It incorporates new modules such as C3k2, SPPF, and C2PSA, and offers five model variants (n/s/m/l/x) for tasks of different scales. Figure 1 shows the basic network framework of YOLO11:

YOLO11 continues the classic three-part architecture of the YOLO series - Backbone, Neck, and Head. Its core innovation lies in the deep optimization of traditional modules. The Backbone section adopts the C2PSA module, integrating channel and spatial attention mechanisms to enhance feature selection capabilities, and introduces the dynamically reparameterized C3k2 module, which uses multi-branch convolution during training to improve feature extraction capabilities while merging into a single 3×3 convolution during inference to maintain efficiency. At the end, an improved SPPF pyramid pooling is used to fuse multi-scale context information. The Neck section is based on the Path Aggregation Network (PAN) structure and uses the C3k2 module to collaboratively optimize the up/down sampling process: in the up-sampling stage, deep semantic features are fused with shallow detail features to enhance the detection of small targets, and in the down-sampling stage, semantic information transmission is reinforced in reverse, forming an adaptive multi-scale feature flow. The Head section adopts a decoupled design to separate classification and

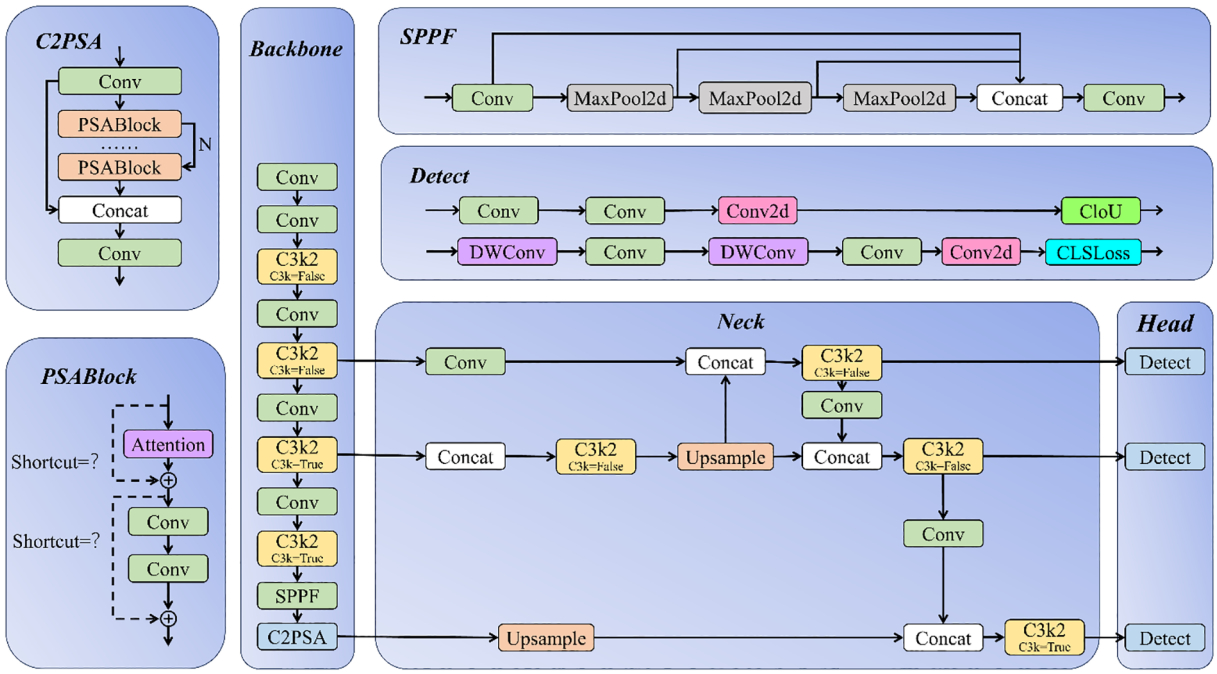

FIGURE 1
Network structure of the YOLO11 model.

localization tasks and innovatively embeds depthwise separable convolution to replace standard convolution layers, significantly reducing computational costs while maintaining accuracy. Compared to YOLOv8, YOLO11, through the three technical breakthroughs of C3k2 reparameterization, C2PSA attention mechanism, and depthwise separable convolution, provides a more robust basic architecture for complex underwater scenarios (such as low-contrast targets in turbid water), which also serves as the key foundation for the dynamic optimization of the DyAqua-YOLO model.

The detection model proposed in this study is based on the YOLO11 + ASF-YOLO architecture. Figure 2 shows the improved network structure diagram:

2.2 Module improvements

2.2.1 DyASF network

ASF-YOLO (Kang et al., 2024) is a new model based on YOLO, proposed by the research team from Monash University Malaysia in 2024. ASF-YOLO integrates the attention scale sequence into the YOLO framework, significantly improving the detection and segmentation performance in scenarios with small and dense targets. Its main structure diagram is shown in Figure 3:

The ASF-YOLO framework is mainly composed of the following components:

1. The SSFF module (Scale Sequence Feature Fusion)

This is used to enhance the ability to extract multi-scale information. It normalizes the feature maps of different scales (such as P3, P4, P5) to the same size and resizes them through nearest neighbor interpolation (Equations 1, 2):

$$\tilde{F}_{P4} = \text{Upsample}_{\text{nearest}}\left(F_{P4}, (H_{P3}, W_{P3})\right) \tag{1}$$

$$\tilde{F}_{P5} = \text{Upsample}_{\text{nearest}}\left(F_{P5}, (H_{P3}, W_{P3})\right) \tag{2}$$

Using 3D convolution combined with multi-scale features (Equations 3, 4):

$$V = \text{Concat}(F_{P3}, \tilde{F}_{P4}, \tilde{F}_{P5}) \in \mathbb{R}^{3\times H\times W\times C} \tag{3}$$

$$F_{\text{SSFF}} = \text{Conv3D}_{3\times3\times3}(V) \tag{4}$$

This module can effectively integrate feature information at different scales, thereby improving the detection and segmentation performance for small targets.

2. TFE Module (Triple Feature Encoder)

It enhances the detection capability for small targets. By concatenating feature maps of three sizes (large, medium, and small), it performs maximum pooling and average pooling downsampling on the large-scale features (Equation 5):

$$F'_{\text{large}} = \frac{1}{2}\left(\text{MaxPool}\left(F_{\text{large}}\right) + \text{AvgPool}\left(F_{\text{large}}\right)\right) \tag{5}$$

Perform nearest-neighbor interpolation upsampling on small-scale features to retain high-resolution features and prevent the loss of small target features (Equation 6):

$$F'_{\text{small}} = \text{Upsample}_{\text{nearest}}\left(F_{\text{small}}, (H_{\text{med}}, W_{\text{med}})\right) \tag{6}$$

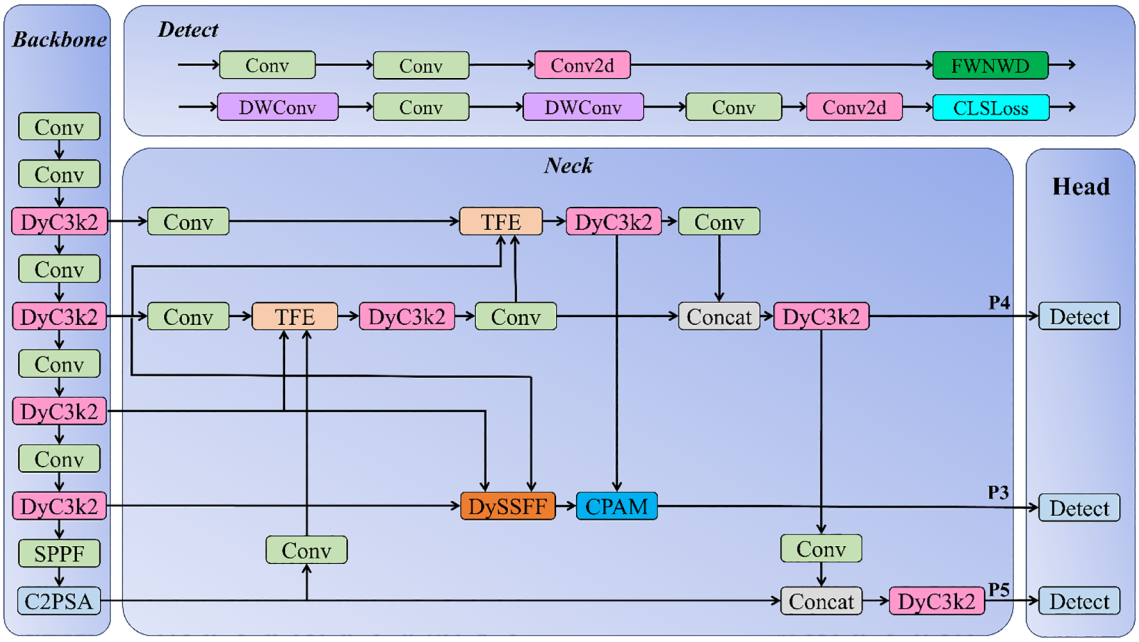

FIGURE 2
Network structure of DyAqua-YOLO model.

Finally, concatenate the multi-resolution features (Equation 7):

$$F_{\text{TFE}} = \text{Concat}\left(F'_{\text{large}}, F_{\text{medium}}, F'_{\text{small}}\right) \tag{7}$$

3. CPAM Mechanism (Channel and Position Attention Mechanism)

The CPAM attention mechanism integrates the feature information of the SSFF and TFE modules. Through the channel attention network and the position attention network, it focuses on the information-rich channels and the position information related to small targets respectively, thereby improving the detection performance. This mechanism can further extract the feature information and enhance the detection effect for small targets. The formula is as follows:

Enhancing key features through channel attention (Equations 8, 9):

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_{\text{TFE}}(i, j, c), w_c = \sigma(\text{Conv1}D_k(z)) \tag{8}$$

$$F_{\text{channel}} = w_c \odot F_{\text{TFE}} \tag{9}$$

$$F_{\text{combined}} = F_{\text{channel}} + F_{\text{SSFF}} \tag{10}$$

Precisely locate through positional attention (Equations 10–12):

$$p_w = \frac{1}{H} \sum_{j=1}^{H} F_{\text{combined}}(:, j),\ p_h = \frac{1}{W} \sum_{i=1}^{W} F_{\text{combined}}(i, :) \tag{11}$$

$$s_{\text{w}}, s_{\text{h}} = \text{Split}(\text{Conv}_{1\times1}(\text{Concat}(p_w, p_h))) \tag{12}$$

Final output fused features (Equation 13):

$$F_{\text{CPAM}} = s_{\text{w}} \otimes s_{\text{h}} \otimes F_{\text{combined}} \tag{13}$$

Based on the original ASF network, we replaced the traditional upsampling operation in the SSFF module of ASF-YOLO with the DySample dynamic upsampling method, obtaining DySSFF (Liu et al., 2023). DySample generates offsets related to the upsampling scale through a linear layer, superimposes them onto the original sampling grid to construct a dynamic sampling set, and implements feature reconstruction based on bilinear interpolation. Compared with traditional upsampling methods, this method adopts a grouped upsampling strategy to reduce interference between channels, and introduces a learnable dynamic range factor to adaptively adjust the intensity of the offset, enhancing feature representation capability and spatial adaptability while maintaining computational efficiency, and improving feature expression.

2.2.2 DyC3k2 module

The traditional C3k2 module employs fixed-parameter convolution kernels, which are difficult to adapt to the dynamic changes of light intensity and the complexity of local features in underwater environments (such as turbid water body false targets and low-contrast biological objects), resulting in limited feature expression capabilities. Figure 4 shows the structure diagram of the C3k2 module. When C3k=False, C3k2 is equivalent to C2f. In the underwater target recognition task, in order to enhance the model's feature extraction ability and adaptability to complex marine environments, we introduced the DynamicConv dynamic convolution mechanism (Han et al., 2024) on the basis of the C3k2 module and designed the DyC3k2 module.

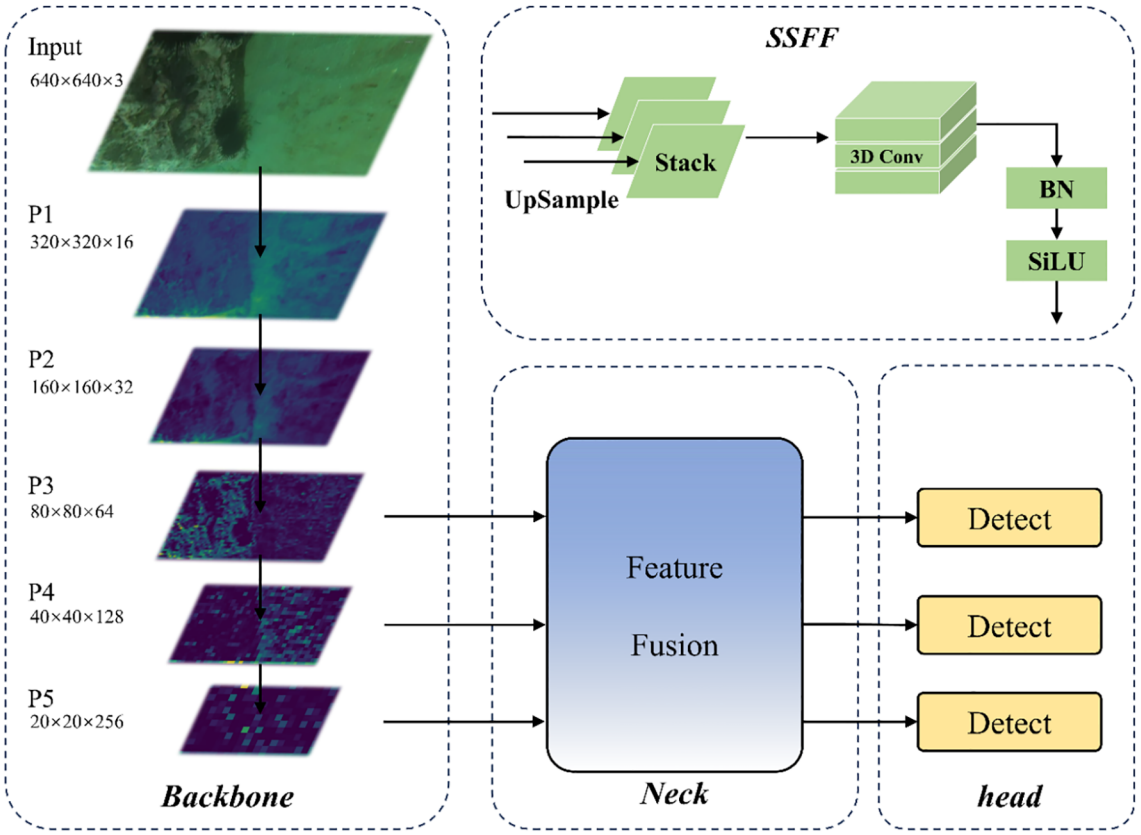

FIGURE 3
Network structure of the ASF-YOLO model.

DynamicConv is a dynamic convolution mechanism that dynamically generates convolution kernel parameters, enabling the model to adaptively adjust the feature extraction method based on the characteristics of the input data. It significantly increases the model's parameters while improving its performance while maintaining low computational complexity (FLOPs). The core idea is to introduce multiple expert convolution kernels and generate dynamic coefficients through a dynamic coefficient generator based on the characteristics of the input samples, which are then used to weight and fuse the weights of these expert convolution kernels and applied to the input feature map. The calculation formula of DynamicConv is as follows:

The input feature map X is obtained as a vector after global average pooling, and then coefficients are generated through two layers of multi-layer perceptrons (MLPs) and an activation function for generating probability distributions (Softmax) (Equation 14):

$$\alpha = \text{softmax}(\text{MLP}(\text{Pool}(X))) \tag{14}$$

The generated weight coefficients are multiplied by the corresponding expert convolution kernels respectively and then summed to obtain the dynamic convolution kernel with a total of M experts (Equation 15):

$$W' = \sum_{i=1}^{M} \alpha_i W_i \tag{15}$$

Convolve the input feature map X with the generated dynamic convolution and obtain the output Y (Equation 16):

$$Y = X * W' \tag{16}$$

2.2.3 FWNWD module

IoU, also known as the Intersection over Union (IoU) (Zhou et al., 2019), is an indicator widely used in the object detection task to measure the similarity between the detected results and the true annotations. In the object detection task, bounding boxes (Bounding Box, Bbox) are usually parameterized by their center point coordinates, width, and height. Specifically, the true bounding box can be represented as $B_{gt} = (x_{gt}, y_{gt}, w_{gt}, h_{gt})$, and the predicted bounding box can be represented as $B = (x, y, w, h)$. The IoU between them is defined as the ratio of their intersection area to their union area, and its mathematical expression is as follows (Equation 17):

$$\text{IoU}(B_{gt}, B) = \frac{\text{Area of overlap } (B_{gt}, B)}{\text{Area of union } (B_{gt}, B)} \tag{17}$$

This indicator quantifies the degree of overlap between the predicted bounding box and the true annotated bounding box. Figure 5 is a geometric illustration of IoU.

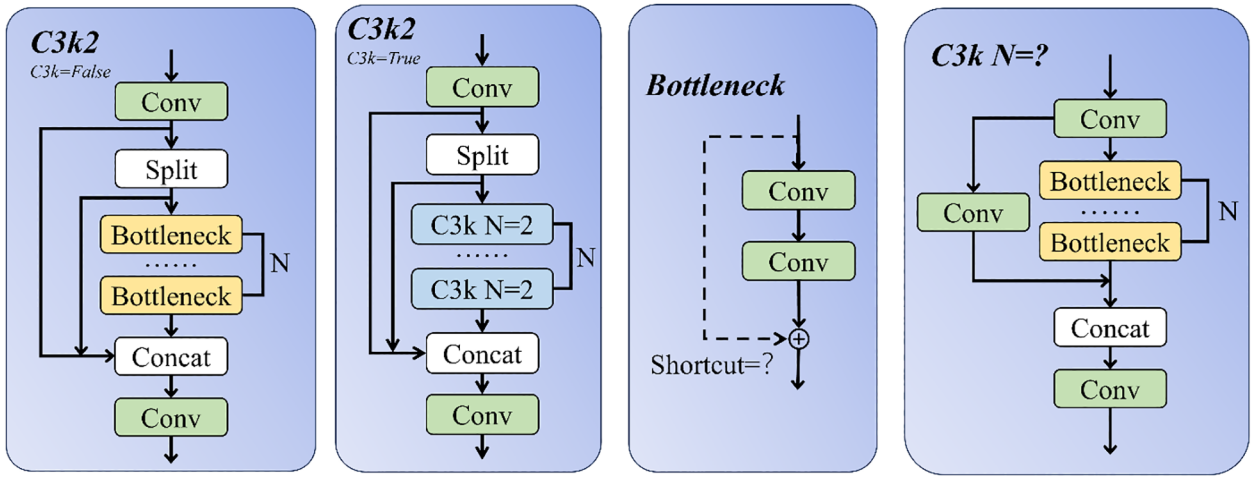

FIGURE 4
Structure diagram of C3k2 module.

In the basic model of YOLO11, the CIoU (Complete Intersection over Union) loss function (Zheng et al., 2020) is used. It integrates the overlapping area, center point distance, and aspect ratio constraints to construct a multi-dimensional geometric supervision mechanism. It introduces a normalized center distance penalty term and a width-to-height similarity measurement term based on IoU, and dynamically balances the geometric feature optimization weights to effectively alleviate the problems of aspect ratio distortion and center offset in bounding box regression. Compared with GIoU (Rezatofighi et al., 2019), it significantly improves the positioning accuracy and accelerates the model convergence.

The complex degradation characteristics of underwater images pose unique challenges for optimizing object detection: gradient interference from low-quality samples, optimization bias caused by uneven distribution of easy and hard samples, and missed detections due to the high sensitivity of small objects to positional deviations. Existing CIoU loss functions exhibit certain limitations in addressing these challenges. To this end, we propose a unified loss function framework named FWNWD (Focaler-Wise Normalized Wasserstein Distance), whose core design concept is to provide an adaptive, multi-objective optimization solution for underwater scenarios through the synergistic integration of three advanced loss mechanisms. The formula proposed in this study for FWNWD is as follows (Equations 18, 19):

$$L_{\mathrm{FWNWD}} = R \times L_{\mathrm{FWIoU}} + (1 - R) \times L_{\mathrm{NWD}} \tag{18}$$

$$L_{\mathrm{FWIoU}} = \mathrm{r} \times R_{\mathrm{WIoU}} \times (1 - \mathrm{IoU}^{\mathrm{focaler}}) \tag{19}$$

Here, L_{FWIoU} is the Focaler-WIoU term that integrates dynamic gradient modulation and sample weighting, L_{NWD} denotes the Normalized Wasserstein Distance (Wang et al., 2021, preprint) that enhances sensitivity to small objects, r and R_{WIoU} are the dynamic focusing coefficient and distance attention term from WIoUv3 (Tong et al., 2023, preprint), $\mathrm{IoU}^{\mathrm{focaler}}$ is the interval-mapped IoU from Focaler-IoU (Zhang and Zhang, 2024, preprint). The configuration of all hyperparameters related to the loss function is provided in Table 1 of Section 3.2.

The construction of FWNWD is based on three core components, with its design motivation stemming from addressing specific optimization challenges in underwater detection tasks: First, to suppress harmful gradient interference caused by low-quality samples, we introduce a dynamic gradient modulation mechanism based on WIoUv3 (Tong et al., 2023, preprint). This mechanism enables adaptive evaluation of sample quality and gradient redistribution by constructing an outlier metric and a dynamic focusing coefficient. By integrating its dynamic focusing coefficient r, FWNWD can automatically identify blurred samples and outliers in underwater environments, effectively enhancing training stability. Second, to tackle the optimization bias caused by uneven distribution of easy and hard samples in underwater scenarios, we adopt the easy-hard sample

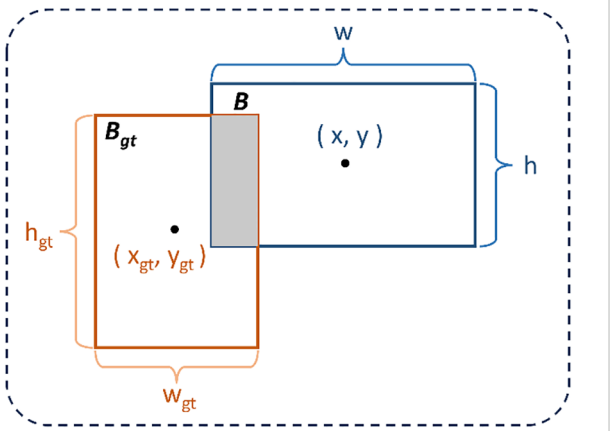

FIGURE 5
Geometric illustration of intersection over union (IoU).

TABLE 1 Hyperparameter configuration for the FWNWD loss function.

Component	Hyperparameter	Description	Value	Rationale
Focaler-IoU	d	Lower confidence limit	0	Recommended value from (Zhang and Zhang, 2024, preprint)
	u	Upper confidence limit	0.95	Recommended value from (Zhang and Zhang, 2024, preprint)
NWD	C	A constant related to the dataset	12.8	Recommended value from (Wang et al., 2021, preprint)
FWNWD	R	Weight balancing L_{FWIoU} and L_{NWD}	0.5	Determined by ablation study (Sec. 3.4.2)

balancing strategy of Focaler-IoU (Zhang and Zhang, 2024, preprint). This strategy redefines the IoU loss function by establishing confidence upper and lower bounds, enabling the model to perform differentiated learning based on sample difficulty. We construct the $IoU^{focaler}$ term using its interval mapping method to achieve targeted learning for hard samples, mitigating the model's tendency to overfit easy samples. Finally, to address the sensitivity to positional deviations in small object detection, we incorporate the NWD-based small object sensitivity metric (Wang et al., 2021, preprint). This method models bounding boxes as Gaussian distributions and computes the Wasserstein distance, constructing a metric more robust to minor positional variations. By introducing the L_{NWD} term, we significantly alleviate the gradient vanishing problem for small objects during training, thereby markedly improving the recall rate of small objects.

The innovation of FWNWD lies in its systematic synergistic architecture design: the aforementioned three technologies are not simply stacked but form a hierarchical optimization framework. The dynamic mechanism of WIoUv3 provides the foundational framework for gradient regulation in the loss function, Focaler-IoU enables differentiated learning for easy and hard samples based on this foundation, and NWD specifically ensures optimization efficiency for small objects. The three components form an organic whole, systematically resolving the complex problem of coexisting multiple types of optimization conflicts in underwater object detection. As shown in the ablation experiments (Section 3.4.2), this combined loss function demonstrates significant advantages in small object recall while comprehensively improving model accuracy.

3 Experiment and discussion

3.1 Dataset

This study utilized the DUO (Dense Underwater Objects) public dataset (Liu et al., 2021a). The DUO dataset was systematically integrated from mainstream datasets such as the URPC (Underwater Robot Picking Challenge) series (URPC, n.d.) and the UDD (Underwater Detection Dataset) (Liu et al., 2021b), and image duplicate removal processing was carried out using the Perceptual Hashing Algorithm (PHash). Eventually, a standardized dataset consisting of 7782 images was formed, including 6671 training images and 1111 test images. This effectively addressed the issues of missing test set annotations, high image redundancy, and varying annotation quality in the URPC series datasets and the UDD dataset, providing a benchmark platform for underwater object detection algorithm research. Figure 6 shows representative images from the DUO dataset.

Compared to the existing datasets, DUO demonstrates three core advantages:

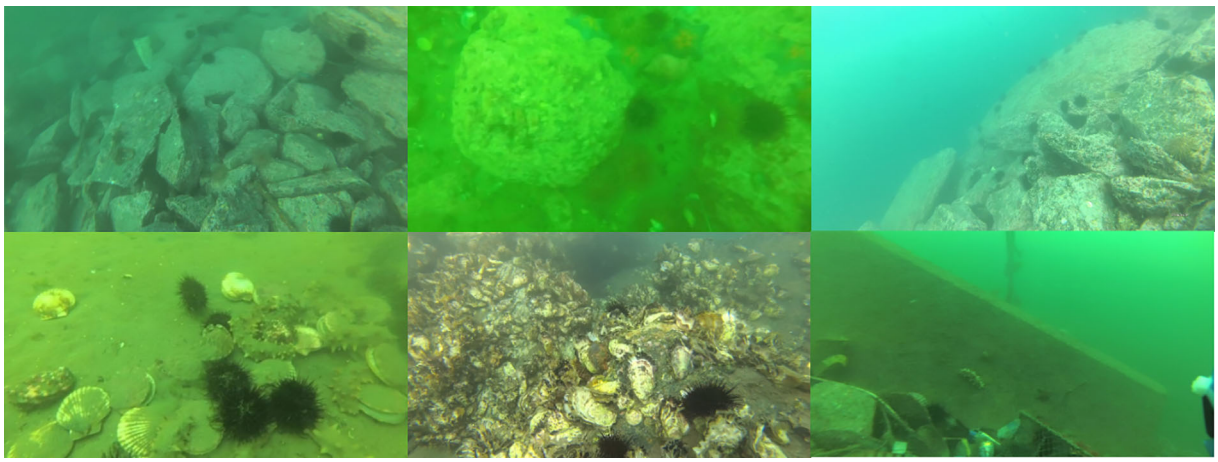

FIGURE 6
Samples of the DUO dataset.

1. Ecological representativeness: It contains 74,515 labeled instances, covering four key species (sea cucumber 10.6%, sea urchin 67.3%, scallop 2.6%, sea star 19.5%), reflecting both the natural long-tail distribution and diverse underwater environmental conditions.
2. Small target dominance: Over 83% of the targets occupy an image area of 0.3 - 1.5%, accurately reproducing the essential challenge of detecting marine organisms in high-resolution underwater images.
3. Dense instance distribution: Each image contains 5–15 organisms (mean: 9.57 ± 3.24), with an instance density that is higher than that of traditional datasets, more realistically simulating the actual detection scenarios.

3.2 Experimental environment and parameter configuration

The experimental environment was set up on the Windows 11 operating system, which has a NVIDIA GeForce RTX 4060 graphics card with a memory size of 8GB and a processor of Intel® Core™ i9-14900HX. The used YOLO framework is Ultralytics 8.3.9, the deep learning framework is torch-2.6.0+cu124, and the development environment is Python 3.11.12.

The training parameter configuration is shown in Table 2:

The rest are set to default.

Furthermore, the proposed FWNWD loss function integrates several advanced mechanisms, whose corresponding hyperparameter settings are crucial to the model's performance. Table 1 details the key hyperparameters involved in FWNWD and its components, including their descriptions, configured values, and the rationale behind their selection. These parameters were either adopted from the default recommendations in their original publications or determined empirically through ablations on the validation split of the underwater DUO dataset to ensure optimal adaptability for our specific task.

3.3 Algorithm evaluation indicators

The performance of the target recognition algorithm needs to be systematically evaluated through multiple quantitative indicators. This section, based on three core dimensions of detection accuracy, robustness, and computational efficiency, elaborates on the definitions, calculation methods, and scientific significance of the mainstream evaluation indicators.

TABLE 2 Training hyper-parameters on DUO datasets.

Config	Parameter
Input image size	640×640
Epochs	100
Batch size	16
Start learning rate	0.01
Optimizer	SGD

3.3.1 Precision and recall

Based on the confusion matrix, precision measures the proportion of correct predictions in the detection results, and recall reflects the proportion of true targets that are correctly detected (Equations 20, 21):

$$Precision = \frac{TP}{TP+FP} \tag{20}$$

$$Recall = \frac{TP}{TP+FN} \tag{21}$$

TP: True Positive count, which requires both correct classification and IoU ≥ threshold;

TN: True Negative count;

FP: False Positive count, representing false detection;

FN: False Negative count, representing missed detection.

3.3.2 Average precision

AP calculates the area under the Precision-Recall curve (Area Under Curve, AUC) to comprehensively evaluate the classification and localization performance of the model at different recall (Equation 22):

$$AP = \int_0^1 P(R)dR \tag{22}$$

3.3.3 Mean average precision

mAP is the arithmetic mean of the AP values for all categories and is the core comprehensive indicator of object detection. Its derivative forms include (Equation 23):

$$mAP = \frac{\sum_{i=1}^{M} AP_i}{M} \tag{23}$$

mAP@0.5: This is a performance metric when the IoU threshold is 0.5, applicable to scenarios with loose positioning.

mAP@[0.5:0.95]: This is a standard metric for the COCO dataset, calculating the average mAP value for IoU thresholds ranging from 0.5 to 0.95 (with a step size of 0.05), to evaluate the model's robustness in terms of positioning accuracy.

3.3.4 Algorithm complexity metric

Parameter quantity: This refers to the total sum of all parameters that need to be learned in the model, which includes but is not limited to the weights of convolutional layers, the weights of fully connected layers, and bias terms, etc. The parameter quantity directly relates to the storage requirements of the model and the memory consumption during training. The formula for calculating the parameter quantity is as follows (Equation 24):

$$Param = (K_h \times K_w \times C_{in}) \times C_{out} + \text{bias} \tag{24}$$

Here, C_{in} represents the number of input channels, C_{out} represents the number of output channels, and $K_h \times K_w$

represents the size of the convolution kernel. $(K_h \times K_w \times C_{in}) \times C_{out}$ represents the weight parameters of the convolution kernel. bias represents the parameter of the bias term, which is a vector of size C_{out}.

FLOPs: It measures the number of floating-point operations performed by the model during forward propagation, including addition, subtraction, multiplication, and division, etc. It is an indicator for evaluating the computational cost of the model during operation.

The formula for calculating FLOPs is as follows (Equation 25):

$$FLOPs = 2 \times (K_h \times K_w \times C_{in} \times H_{out} \times W_{out} \times C_{out}) \quad (25)$$

Here, H_{out} and W_{out} represent the size of the output feature map.

3.4 Experimental comparison and evaluation

3.4.1 Comparison experiment of backbone network models

To comprehensively evaluate the performance of the DyAqua-YOLO model proposed in this study in the underwater object detection task, a comparison experiment was conducted on the DUO dataset using the latest benchmark models of the YOLO series (YOLOv5/v6/v8/v10/11/v12). As shown in Table 3:

The training accuracy comparison chart of different YOLO models is shown in Figure 7. It can be seen that the accuracy indicators of the YOLO11 model framework are close but slightly lower than those of the best-performing benchmark model YOLOv8. The FLOPs are 6.3G, which is close but slightly lower than that of the best-performing benchmark model YOLOv5n (5.8G). However, DyAqua-YOLO is comprehensively ahead with 0.918 mAP@0.5 and 0.722 mAP@[0.5:0.95]. Compared to the benchmark model (YOLO11n), it has improved by 3.9% and 3.7% respectively, proving the superiority of its dynamic feature fusion and adaptive convolution mechanism for underwater targets. Additionally, DyAqua-YOLO has a particularly significant improvement in recall for small targets (Recall: 0.841 *vs* YOLOv8n 0.810), attributed to its dynamic upsampling and optimized loss function mechanism, confirming its adaptability to underwater dense small targets. While maintaining high accuracy, DyAqua-YOLO also has efficiency balance. The parameter size (4.43M) is 65.1% higher than that of YOLOv8n, but the FLOPs (7.5G) only increase by 10.3%, demonstrating the computational benefits of the dynamic architecture.

Due to the lack of publicly available code for direct comparison with several recent underwater-specific models (e.g., MarineYOLO, RG-YOLO), we evaluate the efficacy of our proposed method by comparing its performance against the comprehensive benchmark of general-purpose detectors established in the foundational DUO publication (Liu et al., 2021a). As detailed in Table 3, our model achieves a mAP@0.5 of 91.8%, which significantly outperforms the best result (RepPoints at 80.2% mAP@0.5) reported in the original benchmark. Furthermore, our model achieves a real-time inference speed of 21 FPS (Orange Pi AIpro, 8TOPS NPU, 16GB RAM). This speed is approximately 3x faster than the fastest model benchmarked (FSAF, 7.4FPS, Jetson AGX Xavier; Liu et al., 2021a). Given the marked improvements in both accuracy and efficiency over the established benchmark performance on the DUO dataset, we consider that our method demonstrates highly competitive performance.

3.4.2 Ablation experiment

To validate the effectiveness of each proposed component and determine the optimal hyperparameter for the FWNWD loss, we conduct extensive ablation studies based on the YOLO11n baseline. Firstly, we analyze the impact of the balancing coefficient R in the FWNWD loss. R is designed to trade off the contributions between the L_{FWIoU} term and the L_{NWD} term. We evaluate the model performance on the validation set with different values of R, and the results are summarized in Table 4.

With a small R value (e.g., $R = 0.1$), the NWD term dominates the optimization, and the model achieves peak performance in both Recall and *mAP*@0.5 (0.849 and 92.2%, respectively). This confirms the exceptional effectiveness of NWD in reducing missed detections of small objects and improving detection coverage. As the value of R increases, the influence of the FWIoU term becomes more pronounced. We observe that the model's performance on the stricter and more comprehensive evaluation metric, *mAP*@

TABLE 3 Comparative experimental results of different YOLO models on the DUO dataset.

Model	*Precision*	*Recall*	*mAP*@0.5	*mAP*@[0.5:0.95]	*Param*(M)	*FLOPs*(G)
YOLOv5n	0.865	0.789	0.872	0.671	2.182	5.8
YOLOv6n	0.841	0.764	0.847	0.654	4.155	11.5
YOLOv8n	0.869	0.81	0.886	0.694	2.685	6.8
YOLOv10n	0.851	0.798	0.877	0.658	2.266	6.5
YOLO11n	0.864	0.799	0.879	0.685	2.583	6.3
YOLOv12n	0.858	0.771	0.861	0.666	2.509	5.8
Ours	0.896	0.841	0.918	0.722	4.433	7.5

[0.5:0.95], peaks at R = 0.5 (72.2%). This indicates that assigning a higher weight to FWIoU, while sacrificing a small amount of Recall, enhances the overall localization accuracy of the bounding boxes. Thus, R = 0.5 represents the optimal trade-off between the 'quantity' and 'quality' of detections. It ensures that both components of the FWNWD loss function collaborate effectively, rather than being dominated by a single one. Consequently, we select R = 0.5 as the final configuration to prioritize the demanding requirement for precise target localization in high-precision underwater detection tasks.

Following the determination of R, we proceed to the ablation studies of other modules. The adaptive feature fusion module (ASF), dynamic upsampling (DySample), dynamic convolution (DyC3k2), WIoU bounding box loss, Focaler-IoU loss, and normalized Wasserstein distance (NWD) module were gradually introduced. Table 5 shows the experimental groups of the ablation experiment, and Table 6 presents the experimental results of the ablation experiment. Baseline is YOLO11n.

The experimental results reveal the following key findings:

1. The progressive contribution of the backbone network modules

The training accuracy comparison chart of Baseline-Group4 is shown in Figure 8. From this, we can compare and determine the contribution of each backbone network module to the improvement of the accuracy indicators:

- ASF module (Group 1): Increases mAP@0.5 by 0.4% (0.883 *vs* 0.879), but at the cost of a 12.7% increase in FLOPs (7.1G *vs* 6.3G), indicating that it is effective for blurry targets but the computational efficiency needs to be optimized;
- DyASF module (Group 2): Compared to Group 1, it has improved detection accuracy and small target recall, especially mAP@[0.5:0.95] has increased by 1.4% (0.698

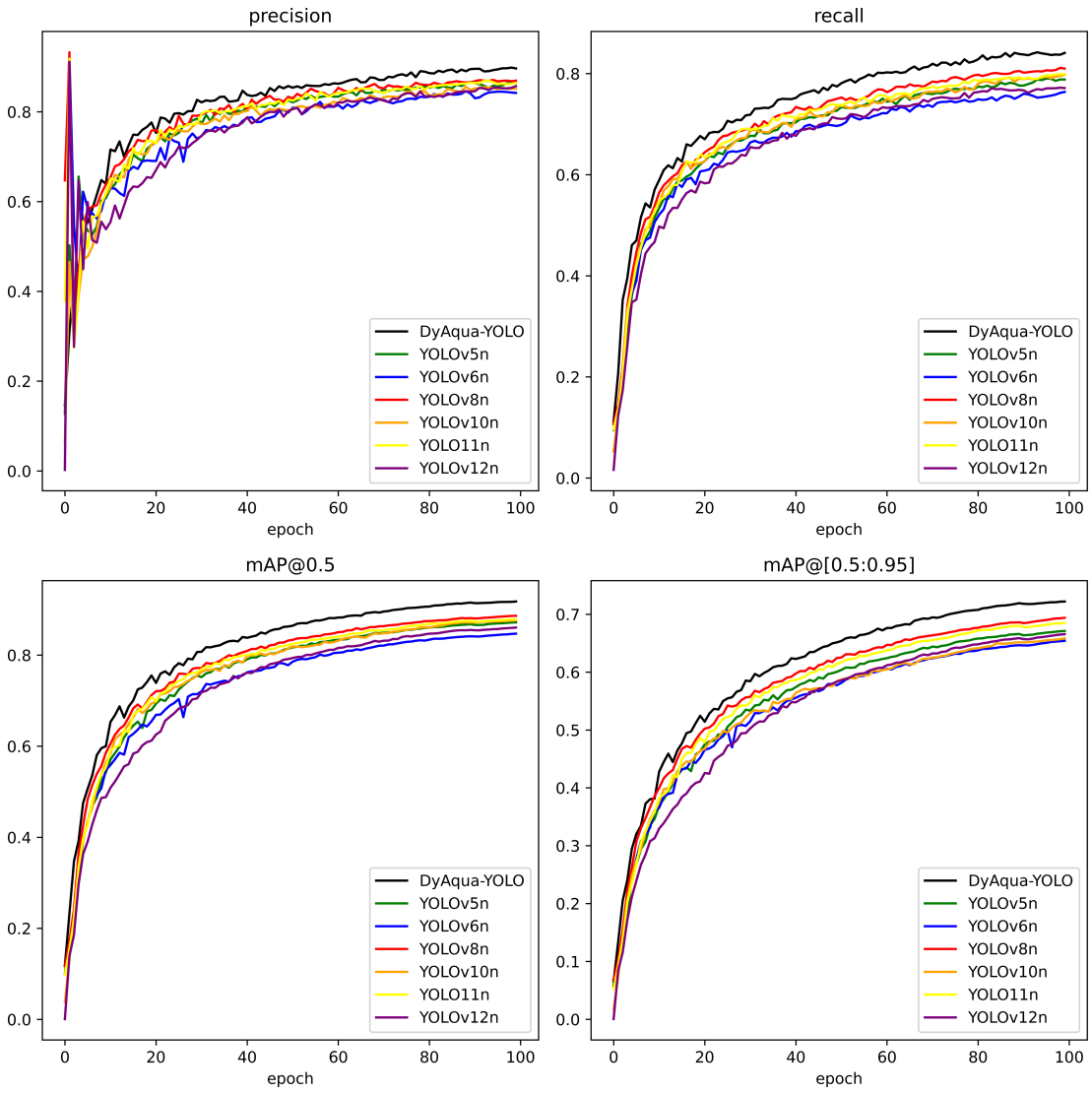

FIGURE 7
Comparison chart of four indicators for different YOLO models.

TABLE 4 Ablation study on the hyperparameter *R*.

R	*Precision*	*Recall*	*mAP*@0.5	*mAP*@[0.5:0.95]
0	0.900	0.842	0.919	0.717
0.1	0.903	0.849	0.922	0.719
0.3	0.890	0.849	0.918	0.719
0.5	0.896	0.841	0.918	0.722
0.7	0.898	0.835	0.917	0.721
0.9	0.899	0.836	0.916	0.720
1	0.889	0.836	0.913	0.719

vs 0.684), indicating that the addition of DySample has made the model have higher detection accuracy, but FLOPs have increased by 12.7%.

- DyC3k2 dynamic convolution combination (Group 3): With only an increase of 0.878M parameters, mAP@0.5 increases by 1.2% (0.891 *vs* 0.879), mAP@[0.5:0.95] increases by 1.7% (0.702 *vs* 0.685), and FLOPs decrease by 3.2% (6.1G *vs* 6.3G), proving that it can adaptively adjust feature extraction to effectively enhance feature representation capability and optimize computational efficiency.

TABLE 5 Grouping of ablation experiments for the DyAqua-YOLO model.

Exp.	ASF	DySample	DyC3k2	WIoU	Focaler-IoU	NWD
Baseline	×	×	×	×	×	×
Group1	✓	×	×	×	×	×
Group2	✓	✓	×	×	×	×
Group3	×	×	✓	×	×	×
Group4	✓	✓	✓	×	×	×
Group5	✓	✓	✓	×	×	✓
Group6	✓	✓	✓	✓	×	×
Group7	✓	✓	✓	✓	✓	×
Group8	✓	✓	✓	✓	×	✓
Ours	✓	✓	✓	✓	✓	✓

The symbol '✓' indicates the inclusion of the corresponding module or component in the model configuration for that specific experiment. Conversely, the symbol 'x' indicates its exclusion.

TABLE 6 Results of ablation experiments for the DyAqua-YOLO model.

Group	*Precision*	*Recall*	*mAP*@0.5	*mAP*@[0.5:0.95]	*Param*(M)	*FLOPs*(G)
Baseline	0.864	0.799	0.879	0.685	2.583	6.3
Group1	0.867	0.803	0.883	0.684	2.675	7.1
Group2	0.869	0.807	0.887	0.698	3.000	7.7
Group3	0.872	0.809	0.891	0.702	3.461	6.1
Group4	0.886	0.805	0.895	0.711	4.433	7.5
Group5	0.890	0.836	0.913	0.720	4.433	7.5
Group6	0.890	0.833	0.911	0.721	4.433	7.5
Group7	0.889	0.836	0.913	0.719	4.433	7.5
Group8	0.897	0.832	0.915	0.720	4.433	7.5
Ours	0.896	0.841	0.918	0.722	4.433	7.5

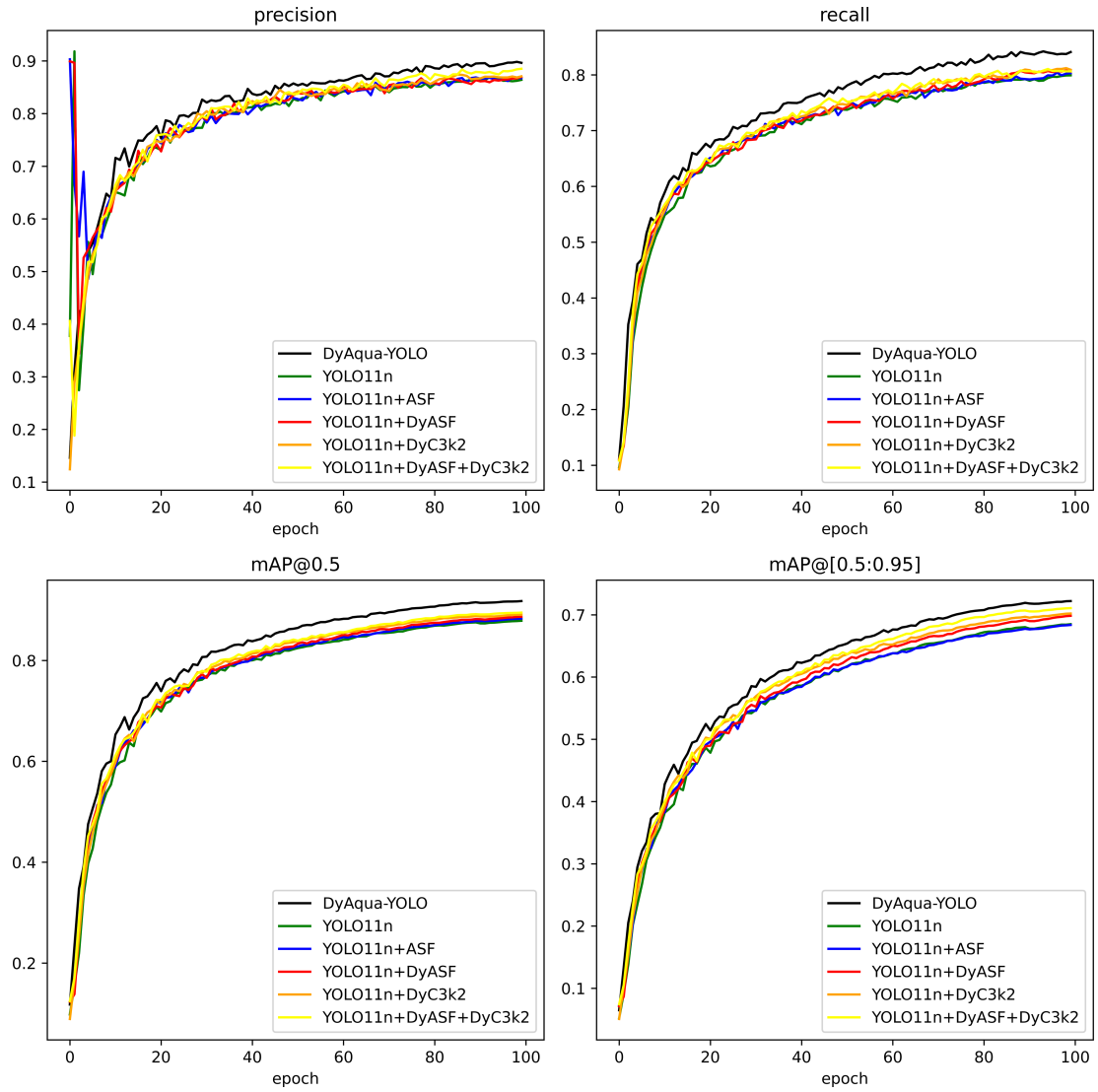

FIGURE 8
Comparative diagram of model ablation experiments based on YOLO11.

- Multi-module combination (Group 4): After adding multiple modules, a combined effect was produced. Compared to the baseline, mAP@0.5 increased by 1.6% (0.895 *vs* 0.879), mAP@[0.5:0.95] increased by 2.6% (0.711 *vs* 0.685), and FLOPs increased by 19% (7.5G *vs* 6.3G).

2. Synergistic effect of the loss function

The training accuracy comparison chart of Group4-Ours is shown in Figure 9, and the training loss comparison chart is shown in Figure 10. It can be seen that the dependent variable in this case is the loss function. From this, we can compare and determine the contribution of our newly proposed FWNWD loss function. Figure 9 is a line chart showing the impact of different loss functions on the model training accuracy, and Figure 10 is a chart showing the influence of different loss functions on the model training loss.

- Compared with Group4, using the NWD module alone (Group5) and the WIoU loss (Group6) can both increase detection accuracy without increasing computational burden. The mAP@0.5 is improved by 1.8% (0.913 *vs* 0.895) and 1.6% (0.911 *vs* 0.895) respectively, and the mAP@[0.5:0.95] is increased by 0.9% (0.720 *vs* 0.711) and 1.0% (0.721 *vs* 0.711) respectively. The recall of small targets has significantly increased by 3.1% and 2.8% respectively, indicating that these two loss functions can effectively improve the positioning accuracy of fuzzy targets and reduce false detection and missed detection rates.
- Compared with WIoU (Group6), Focaler-WIoU (Group7) improves mAP@0.5 by 0.2% without increasing computational burden, but the mAP@[0.5:0.95] decreases by 0.2%. The recall of small targets increases by 0.3%. This shows that Focaler-WIoU has certain efficacy in focusing on difficult samples.

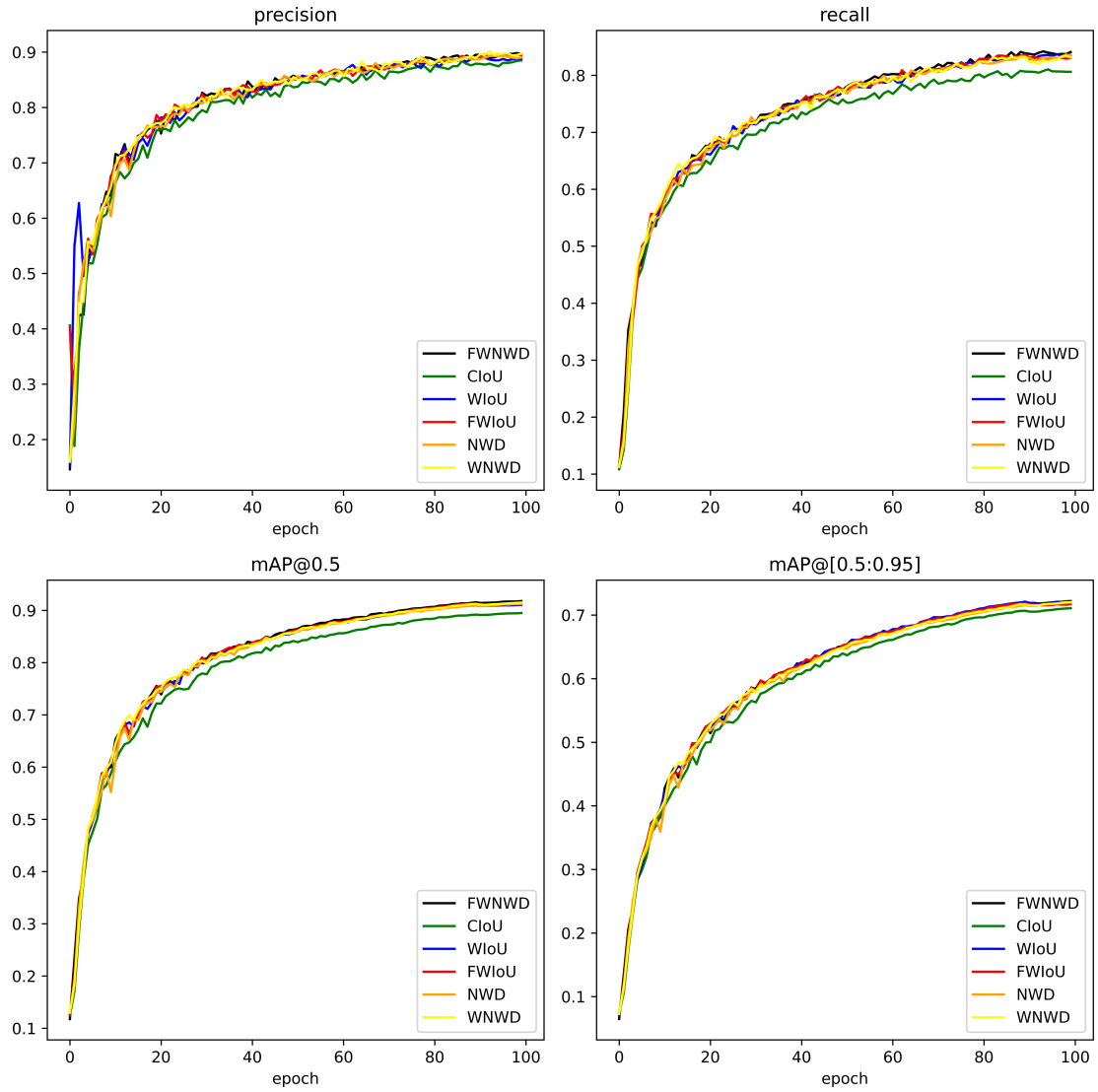

FIGURE 9
The impact of different loss functions on the training accuracy of the model.

- A single loss function is difficult to balance the retrieval of small targets and the positioning accuracy of multi-scale, but when the three are coordinated and complement each other, the model can simultaneously consider the recall and accuracy indicators of small targets and achieve better results. Compared with using the CIoU loss function, after adding the FWNWD loss function module, mAP@0.5 and mAP@[0.5:0.95] increase by 2.3% and 1.1% respectively, and the recall of small targets increases by 3.6%.

3.4.3 Detection effect presentation

To verify the performance superiority of the DyAqua-YOLO model proposed in this study compared to the mainstream models, we selected four representative underwater images from the test set for visual comparison and analysis of the detection results. Figure 11 shows the original images and the prediction results of YOLOv5, YOLOv6, YOLOv8, YOLOv10, YOLO11, YOLOv12, and the DyAqua-YOLO model of this study, corresponding to the eight columns in the figure respectively.

In Figure 11, The ground truths are denoted by purple boxes and labels, while the predictions of our model are denoted by red boxes and labels. Image (a) has an original size of 3840×2160 pixels, with a total of 16 annotated objects. The image features high resolution, partial occlusion of some objects, and generally small target sizes (the smallest occupying 0.076% of the image area, the largest 1.10%). During the training phase, image compression can easily lead to the loss of features in small objects, increasing the difficulty of detection. Compared to other models, DyAqua-YOLO demonstrates excellent small object detection capability, with no missed detections and only 2 false positives. Image (b) has an original size of 1920×1080 pixels. Due to turbid water and motion blur, the image is blurred, posing significant challenges for object

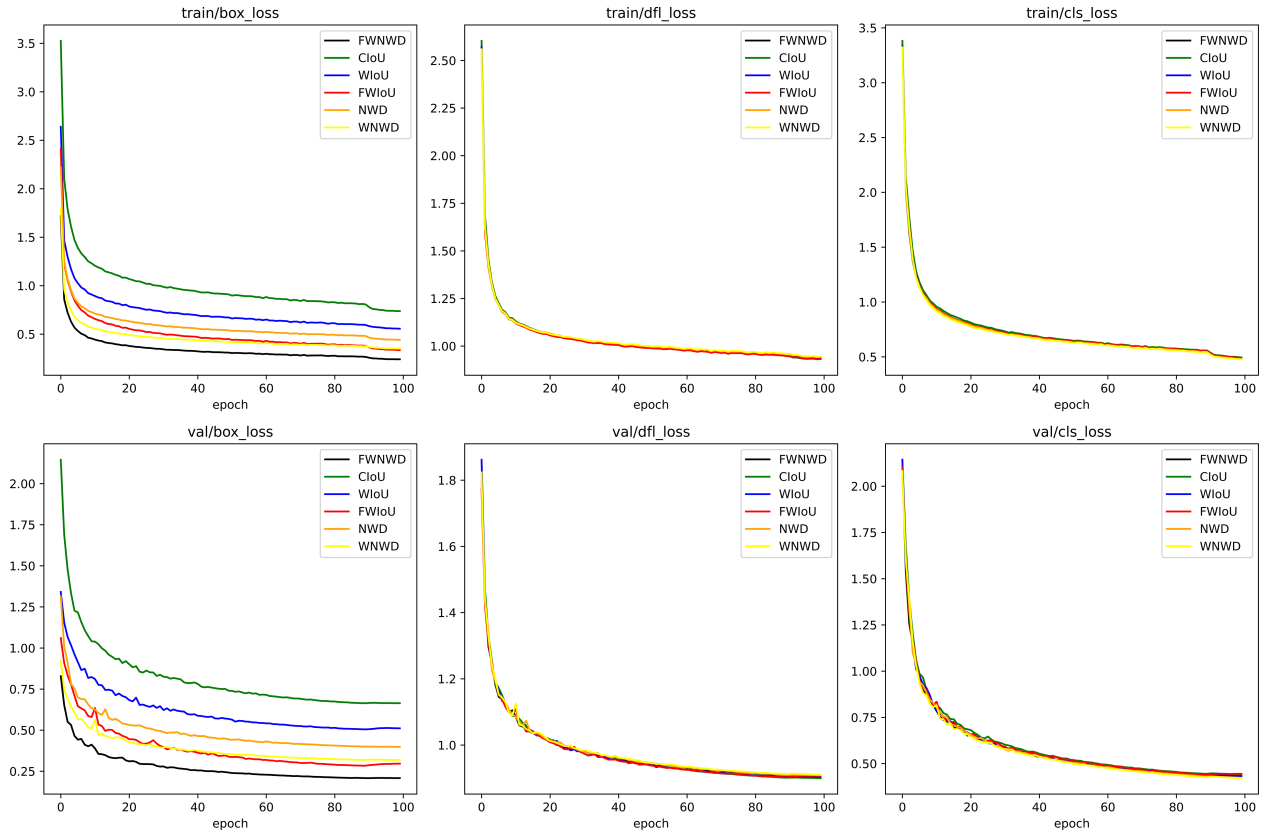

FIGURE 10
The impact of different loss functions on the training loss of the model.

detection. Except for DyAqua-YOLO and YOLOv5, all other models exhibit missed or false detections. Further comparison between DyAqua-YOLO and YOLOv5 shows that the former demonstrates superior localization accuracy. Image (c) has an original size of 720×405 pixels, containing 30 annotated objects with dense distribution, occlusion, and overlapping. DyAqua-YOLO successfully detects 29 objects, with only 1 missed detection. Although its perception mechanism for low-quality samples introduces a small number of false positives, the overall detection performance is significantly better than that of other compared models. Image (d) has an original size of 3840×2160 pixels, with a complex background and two occluded objects. Models including YOLOv5, YOLOv8, YOLOv10, YOLOv11, and YOLOv12 are affected by the complex background and occlusion, resulting in missed or false detections. Although YOLOv6 is not interfered by the background, it still misses one occluded object. DyAqua-YOLO demonstrates strong anti-interference capability in this image, delivering robust detection results.

Overall, the DyAqua-YOLO model demonstrates superior comprehensive performance across a variety of challenging underwater scenarios. It exhibits strong anti-interference capability, effectively suppressing false positives and missed detections caused by complex backgrounds and occlusions. In the presence of dense targets, partial occlusion, or small targets, the model achieves a higher detection rate. Furthermore, compared to other models in the YOLO series (including YOLOv5, v6, v8, v10, v11, and v12), which occasionally suffer from detection box misalignment or localization drift, DyAqua-YOLO produces bounding boxes that are generally more accurate and consistent with the actual targets, indicating more stable performance. These experimental results clearly demonstrate that DyAqua-YOLO holds significant advantages over current mainstream YOLO models when addressing key challenges in underwater object detection, such as occlusion, blurry imagery, target density, and complex background noise.

4 Conclusion and outlook

This study addresses the core algorithmic challenges presented by underwater environments, specifically the detection of blurred, low-contrast, and small targets that result from complex optical degradation. It innovatively proposes the DyAqua-YOLO model based on a dynamically adjustable architecture. The key breakthrough lies not in the individual modules but in their systematic co-design, which creates a synergistic effect greater than the sum of its parts. By deeply integrating the DySSFF and the DyC3k2, it achieves the collaborative dynamic optimization of multi-scale feature representation and convolution kernel weights. Additionally, the designed FWNWD loss function innovatively

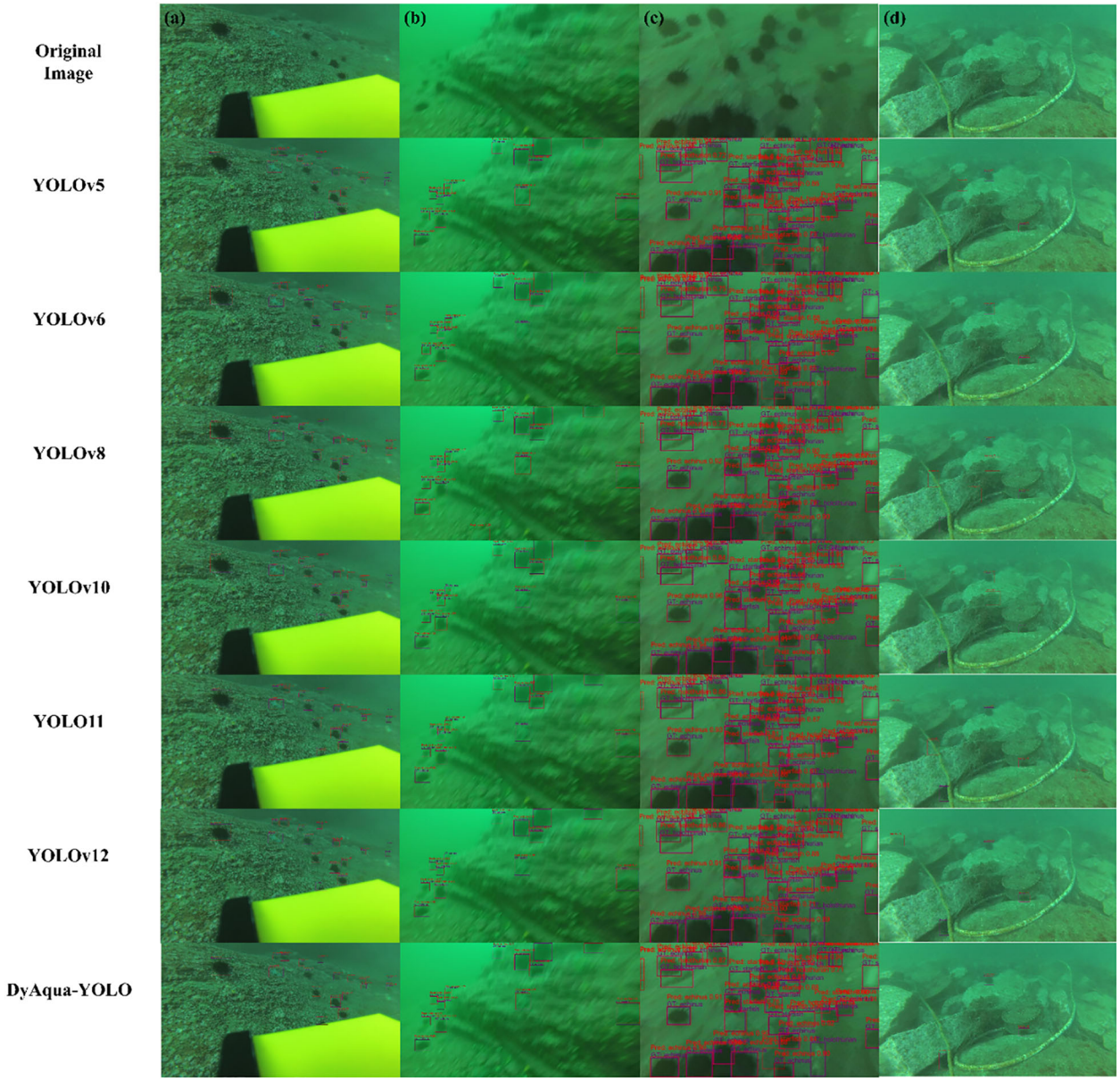

FIGURE 11
Visual comparison of detection results from different models on four challenging underwater scenes. **(a)** high-resolution image with small and partially occluded objects, **(b)** blurred image in turbid water, **(c)** scene with dense and overlapping objects, **(d)** complex background with occlusions.

combines the WIoU dynamic focusing mechanism, Focaler-IoU sample weighting strategy, and NWD small target metric, significantly enhancing the model's robustness for detecting targets in turbid water bodies. Systematic experiments on the DUO dataset demonstrate that DyAqua-YOLO significantly improves underwater object detection performance by 91.8% mAP@0.5 and 72.2% mAP@[0.5:0.95], outperforming the baseline model by 3.9%, and meeting real-time requirements at 21FPS (OrangePi AIpro, 8TOPS NPU, 16GB RAM). Ablation experiments further reveal the cascading gain effect of the dynamic module - DyC3k2 increases the small target recall by 3.6%, while the FWNWD loss effectively resolves the trade-off between difficult sample optimization and micro-object detection.

Although DyAqua-YOLO demonstrates promising performance in underwater object detection, several research directions deserve further exploration to enhance its capabilities and practical applicability. Firstly, to alleviate the information loss caused by image downsampling, we plan to develop adaptive high-resolution processing strategies, such as adaptive image tiling and multi-scale inference mechanisms. These approaches aim to preserve fine-grained features of small objects while maintaining computational efficiency. In addition, underwater image enhancement techniques—including deblurring, contrast enhancement, and color correction—will be investigated to improve input image quality and provide more reliable visual information for detection. Second, we will focus on lightweight and hardware-aware model optimization to facilitate deployment on resource-constrained

embedded platforms. Techniques such as neural architecture search (NAS), quantization, and pruning will be employed to reduce computational and memory overhead without significantly compromising detection accuracy. Furthermore, to address perception challenges in complex underwater optical environments, we intend to construct an acoustic-optical multi-modal perception framework. This system will leverage acoustic imaging to compensate for the lack of visual information in highly turbid water, thereby improving detection robustness in low-visibility conditions. Finally, the dynamic architecture and loss function proposed in this study show potential for generalization beyond underwater detection. We plan to extend their application to other vision tasks such as aerial image analysis and medical image recognition, evaluating their adaptability and effectiveness across diverse domains.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

SL: Formal analysis, Validation, Visualization, Writing – original draft. XP: Funding acquisition, Investigation, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research and/or publication of this article. This work was supported by the National Key Laboratory Foundation (WDZC20245250303), the Independent Innovation Science Fund of National University of Defense Technology (24-ZZCX-JDZ-29),the National Natural Science Foundation of China (52305594), and the Natural Science Foundation of Hunan Province (2024JJ6460).

Conflict of interest

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

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

Amjoud, A. B., and Amrouch, M. (2023). Object detection using deep learning, CNNs and vision transformers: A review. *IEEE Access* 11, 35479–35516. doi: 10.1109/ACCESS.2023.3266093

Chen, X., Fan, C., Shi, J., Wang, H., and Yao, H. (2024). Underwater target detection and embedded deployment based on lightweight YOLO_GN. *J. Supercomput.* 80, 14057–14084. doi: 10.1007/s11227-024-06020-0

Cheng, C., Wang, C., Yang, D., Wen, X., Liu, W., and Zhang, F. (2024). Underwater small target detection based on dynamic convolution and attention mechanism. *Front. Mar. Sci.* 11. doi: 10.3389/fmars.2024.1348883

Dai, J., Li, Y., He, K., and Sun, J. (2016). R-FCN: object detection via region-based fully convolutional networks. *Adv. Neural Inf. Process. Syst.* 379–387. doi: 10.48550/arXiv.1605.06409

Dalal, N., and Triggs, B. (2005). Histograms of oriented gradients for human detection. *Proc. IEEE Comput. Soc Conf. Comput. Vis. Pattern Recognit.* 1, 886–893. doi: 10.1109/cvpr.2005.177

Feng, G., Xiong, Z., Pang, H., Gao, Y., Zhang, Z., Yang, J., et al. (2024). RTL-YOLOV8N: a lightweight model for efficient and accurate underwater target detection. *Fishes* 9, 294. doi: 10.3390/fishes9080294

Girshick, R. (2015).). Fast R-CNN. *Proc. IEEE Int. Conf. Comput. Vis.*, 1440–1448. doi: 10.1109/ICCV.2015.169

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE conf. Comput. Vis. Pattern Recognit.* Piscataway, 580–587. doi: 10.1109/CVPR.2014.81

Han, K., Wang, Y., Guo, J., and Wu, E. (2024). "ParameterNet: parameters are all you need for large-scale visual pretraining of mobile networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* Piscataway, 15751–15761. doi: 10.1109/cvpr52733.2024.01491

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 1904–1916. doi: 10.1109/TPAMI.2015.2389824

Hu, Z., Cheng, L., Yu, S., Xu, P., Zhang, P., Tian, R., et al. (2025). Underwater target detection with high accuracy and speed based on YOLOv10. *J. Mar. Sci. Eng.* 13, 135. doi: 10.3390/jmse13010135

Kang, M., Ting, C., Ting, F. F., and Phan, R. C. (2024). ASF-YOLO: a novel YOLO model with attentional scale sequence fusion for cell instance segmentation. *Image Vis. Comput.* 147, 105057. doi: 10.1016/j.imavis.2024.105057

Khanam, R., and Hussain, M. (2024). *). Yolov11: An overview of the key architectural enhancements*.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., et al. (2016). SSD: single shot multibox detector. *Lect. Notes Comput. Sci.*, 21–37. doi: 10.1007/978-3-319-46448-0_2

Liu, L., Chu, C., Chen, C., and Huang, S. (2024). MarineYOLO: Innovative deep learning method for small target detection in underwater environments. *Alexandria Eng. J.* 104, 423–433. doi: 10.1016/j.aej.2024.07.126

Liu, C., Li, H., Wang, S., Zhu, M., Wang, D., Fan, X., et al. (2021a). "A dataset and benchmark of underwater object detection for robot picking," in *Proc. IEEE Int.*

Conf. Multimed. Expo Workshops Piscataway, 1–6. doi: 10.1109/ICMEW53276.2021.9455997

Liu, B., Liu, Z., Men, S., Li, Y., Ding, Z., He, J., et al. (2020). Underwater hyperspectral imaging technology and its applications for detecting and mapping the seafloor: A review. *Sensors* 20, 4962. doi: 10.3390/s20174962

Liu, W., Lu, H., Fu, H., and Cao, Z. (2023). "Learning to upsample by learning to sample," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.* Piscataway, 6004–6014. doi: 10.1109/iccv51070.2023.00554

Liu, C., Wang, Z., Wang, S., Tang, T., Tao, Y., Yang, C., et al. (2021b). A new dataset, poisson GAN and AquaNet for underwater object grabbing. *IEEE Trans. Circuits Syst. Video Technol.* 32, 2831–2844. doi: 10.1109/tcsvt.2021.3100059

Lu, D., Yi, J., and Wang, J. (2024). Enhanced YOLOV7 for improved underwater target detection. *J. Mar. Sci. Eng.* 12, 1127. doi: 10.3390/jmse12071127

Luo, Y., and Feng, W. (2025). Target detection and image enhancement for underwater environment: Research on improving YOLOv7. *IEEE Access* 13, 34831–34843. doi: 10.1109/ACCESS.2025.3544061

Luo, H., Ruan, H., and Tu, D. (2024). Research on small sample target detection for underwater robot. *Robot. Intell. Autom.* 44, 229–241. doi: 10.1108/ria-07-2023-0090

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). "You only look once: unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* Piscataway, 779–788. doi: 10.1109/CVPR.2016.91

Ren, S., He, K., Girshick, R., and Sun, J. (2017). Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 1137–1149. doi: 10.1109/TPAMI.2016.2577031

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., and Savarese, S. (2019). "Generalized intersection over union: a metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* Piscataway, 658–666. doi: 10.1109/cvpr.2019.00075

Tong, Z., Chen, Y., Xu, Z., and Yu, R. (2023). *Wise-IoU: bounding box regression loss with dynamic focusing mechanism*.

United Nations (2022). *The second world ocean assessment*. Available online at: https://digitallibrary.un.org/record/3978757 (Accessed September 14, 2025).

URPC *Underwater robot professional contest*. Available online at: http://www.urpc.org.cn/index.html (Accessed September 14, 2025).

Viola, P., and Jones, M. (2001). "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* Piscataway, I–I. doi: 10.1109/CVPR.2001.990517

Wang, J., Xu, C., Yang, W., and Yu, L. (2021). *A normalized Gaussian Wasserstein distance for tiny object detection*.

Xu, S., Zhang, M., Song, W., Mei, H., He, Q., and Liotta, A. (2023). A systematic review and analysis of deep learning-based underwater object detection. *Neurocomputing* 527, 204–232. doi: 10.1016/j.neucom.2023.01.056

Zaidi, S. S. A., Ansari, M. S., Aslam, A., Kanwal, N., Asghar, M., and Lee, B. (2022). A survey of modern deep learning based object detection models. *Digit. Signal Process.* 126, 103514. doi: 10.1016/j.dsp.2022.103514

Zhang, H., Dai, C., Chen, C., Zhao, Z., and Lin, M. (2024). One stage multi-scale efficient network for underwater target detection. *Rev. Sci. Instrum.* 95, 065108. doi: 10.1063/5.0206734

Zhang, H., and Zhang, S. (2024). *Focaler-IoU: More Focused Intersection over Union Loss*.

Zheng, K., Liang, H., Zhao, H., Chen, Z., Xie, G., Li, L., et al. (2024). Application and analysis of the MFF-YOLOV7 model in underwater sonar image target detection. *J. Mar. Sci. Eng.* 12, 2326. doi: 10.3390/jmse12122326

Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., and Ren, D. (2020). Distance-IOU loss: faster and better learning for bounding box regression. *Proc. AAAI Conf. Artif. Intell.* 34, 12993–13000. doi: 10.1609/aaai.v34i07.6999

Zheng, Z., and Yu, W. (2024). RG-YOLO: multi-scale feature learning for underwater target detection. *Multimed. Syst.* 31, 26. doi: 10.1007/s00530-024-01617-0

Zhou, D., Fang, J., Song, X., Guan, C., Yin, J., Dai, Y., et al. (2019). Iou loss for 2d/3d object detection. *Proc. Int. Conf. 3D Vis.*, 85–94. doi: 10.1109/3DV.2019.00019