frontiers in Marine Science

TYPE Original Research
PUBLISHED 03 June 2026
DOI 10.3389/fmars.2026.1859313

OPEN ACCESS

EDITED BY
Lin Liu,
Ministry of Natural Resources, China

REVIEWED BY
Weichao Pan,
Shandong Jianzhu University, China
Haodi Zhu,
Zhejiang University, China

*CORRESPONDENCE
Chao Zhang
dyc.zhangchao@gmail.com

RECEIVED 18 April 2026
REVISED 18 May 2026
ACCEPTED 19 May 2026
PUBLISHED 03 June 2026

CITATION
Zhang C, Li X, Wu S, Zhang Z, Cao X, Xia T and Yu W (2026) An underwater dual-modal denoising detection network for illumination fluctuation and dense occlusion scenes.
Front. Mar. Sci. 13:1859313.
doi: 10.3389/fmars.2026.1859313

An underwater dual-modal denoising detection network for illumination fluctuation and dense occlusion scenes

Chao Zhang[1]*, Xingkun Li[2], Shuang Wu[3], Zhongfeng Zhang[3], Xiangyang Cao[1], Tiantian Xia[3] and Wenlong Yu[1]

[1]College of Transport Geography, Shandong Jiaotong University, Jinan, China, [2]School of Physical Sciences, Qingdao University, Qingdao, China, [3]College of Land Resources and Surveying Engineering, Shandong Agriculture and Engineering University, Jinan, China

Background: Underwater object detection is a critical enabling technology for intelligent ocean exploration. However, light scattering and absorption in underwater environments cause significant RGB image degradation and texture loss, leading to weak foreground feature representation and limited detection accuracy in densely occluded scenes.

Methods: To address these challenges, this paper proposes the Underwater Dual-modal Detection Network (UW-DualDet), which incorporates depth information to compensate for RGB degradation and enhance detection performance. Two plug-and-play modules are designed: the Underwater Denoise Feature Fusion (UDFF) module suppresses dual-modal noise and adaptively exploits the complementarity between RGB texture and depth geometry to mitigate single-modality failure under extreme illumination, while the Underwater Feature Enhancement Module (UFEM) improves feature representation and multiscale semantic perception through multi-scale depthwise separable convolutions and a multi-dimensional weighting mechanism. Furthermore, a Gaussian-filterenhanced YOLO26 detection head (D26Head) is introduced to reduce missed and false detections in densely occluded scenarios by balancing positive sample coverage and localization accuracy through a dual-branch optimization strategy, while also suppressing feature noise.

Results: UW-DualDet achieves a mean Average Precision (mAP) of 86.7%, outperforming all compared state-of-the-art methods. It demonstrates superior robustness across three representative scenarios: 88.2% AP_{50} in clear water, 79.6% AP_{50} in turbid environments, and 75.3% AP_{50} in low-light deep water. The model maintains an inference speed of 116 FPS, meeting real-time operational requirements.

Discussion: The proposed method effectively addresses the core challenges of underwater object detection by leveraging multimodal complementarity and task-specific optimization. It provides robust technical support for underwater intelligent operations and ecological monitoring. Future work will focus on improving depth estimation accuracy and extending the model to more extreme underwater environments.

KEYWORDS
deep learning, feature enhancement, multimodal feature fusion, multimodal object detection, underwater object detection

1 Introduction

The ocean is a crucial component of Earth's environment, containing abundant biological resources, mineral reserves, and energy assets Zhao et al. (2025); Cai et al. (2025); Xu et al. (2025); Rao et al. (2025). It also serves as a critical setting for ecological conservation and the autonomous operation of underwater intelligent systems Gao et al. (2024); Hua et al. (2023a); Zhao et al. (2023); Zhang et al. (2024). As a core task in underwater visual perception, object detection enables accurate localization and classification of submerged targets Chen et al. (2024a). It thus provides essential support for applications such as marine resource exploration and ecological monitoring, while promoting the intelligent and autonomous development of ocean exploration Jian et al. (2024); Wang et al. (2023b).

Unlike terrestrial environments, underwater scenes are severely affected by light absorption and scattering Fu et al. (2023); Jin et al. (2022); Han et al. (2025). As a result, acquired RGB images often exhibit significant degradation, including blurred edges, texture loss, reduced contrast, and color distortion. Additionally, underwater illumination varies dramatically, and image quality differs markedly between sunlit near-surface environments and dim deep-water conditions. Dense target distributions, mutual occlusions among marine organisms, and obstructions caused by reefs further complicate feature extraction and detection accuracy. As shown in Figure 1, these factors make conventional terrestrial detection models highly susceptible to feature representation failure and significant performance degradation in underwater environments.

In current underwater object detection research, single-modal methods remain the dominant paradigm. These approaches typically rely on detection frameworks like YOLO and Faster R-CNN, enhancing model adaptability to degraded underwater RGB images through techniques such as image enhancement, channel and spatial attention, and multi-scale feature fusion Song et al. (2023); Guo et al. (2024); Zhang and Gao (2025); Li and Peng (2025). However, they still depend exclusively on RGB information and cannot fundamentally compensate for feature loss caused by scattering and absorption in water. As a result, their detection accuracy and robustness remain limited in complex underwater environments with extreme illumination variations and high turbidity Chen et al. (2022); Wang et al. (2023a, 2025).

Advances in underwater sensing technology have enabled depth images to provide geometric and spatial information that RGB images alone cannot capture Yang et al. (2024). This information complements the texture and semantic cues in RGB data. Accordingly, RGB-depth multimodal fusion has become an important direction in underwater object detection Yang et al. (2025). However, most existing multimodal approaches rely on basic fusion strategies, such as channel concatenation or element-wise addition, without addressing the inherent noise in underwater bimodal data. These methods often amplify noise and introduce feature redundancy during fusion. Moreover, they generally lack task-specific feature enhancement modules and optimized detection heads suited to underwater scenes. This limits their effectiveness in challenging conditions, such as dense occlusion and extreme illumination changes, and prevents the full exploitation of cross-modal complementarity.

To address the limitations of existing underwater object detection methods, such as insufficient feature representation in single-modal models, inefficient multimodal fusion, and poor adaptability to complex scenes, this paper proposes UW-DualDet, an end-to-end RGB-depth bimodal detection network for underwater environments. The network takes synchronized RGB and depth images as input and uses two parallel backbone branches with independent parameters to extract texture-semantic features from the RGB modality and geometric-structural features from the depth modality, respectively. This design preserves the intrinsic characteristics of both modalities, avoiding the loss of complementary information and noise amplification associated with early fusion. At corresponding scales, a dedicated underwater denoising feature fusion module, UDFF, is introduced to hierarchically fuse bimodal features. By suppressing modality-specific noise and dynamically assigning fusion weights via an adaptive gating mechanism, UDFF enhances cross-modal complementarity and mitigates feature degradation under extreme illumination and noisy underwater conditions. The fused multi-scale features are passed to a bidirectional feature pyramid for cross-level information interaction. In addition, an underwater feature enhancement module, UFEM, is incorporated to strengthen fine-grained details and global semantic representations of low-contrast and small-scale targets, improving adaptability to multi-scale underwater objects. Finally, the enhanced features are fed into a dual-branch detection head, D26Head, which integrates two complementary detection paradigms and a denoising interaction unit to jointly optimize classification and localization. This design effectively reduces missed detections and false positives in densely occluded scenarios, enabling accurate end-to-end underwater object detection. The main contributions of this paper are summarized as follows:

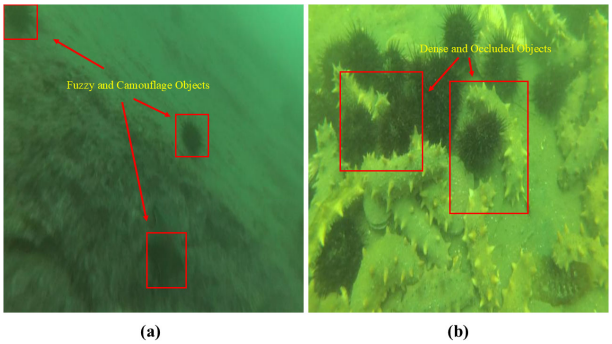

FIGURE 1
Challenges in underwater object detection: **(a)** blurry and camouflage targets; **(b)** dense and occluded targets.

- An end-to-end RGB-depth bimodal object detection network, UW-DualDet, is proposed for complex underwater scenes. The model follows a fully optimized architecture consisting of bimodal parallel feature extraction, multi-scale hierarchical denoising fusion, feature pyramid enhancement, and dualbranch high-precision detection. This architecture addresses the information limitations of traditional single-modal methods and the low fusion efficiency of existing multimodal approaches, providing a new framework for underwater object detection.
- A plug-and-play underwater denoising feature fusion module, UDFF, is designed. UDFF combines a Gaussian denoising unit with an adaptive gating mechanism to dynamically exploit the complementarity between RGB texture cues and depth geometric information while suppressing modality-specific noise. This design alleviates single-modal feature failure under extreme illumination and reduces noise amplification during fusion.
- An underwater feature enhancement module, UFEM, is introduced. Built on multi-scale depthwise separable convolutions and a multi-dimensional weighting mechanism, UFEM enhances fine details and global semantic representations for low-contrast and small-scale targets with minimal computational overhead. It compensates for feature loss caused by scattering and absorption, improving the model's adaptability to targets of different scales.
- A dual-branch detection head, D26Head, is developed. It integrates one-to-one and one-to-many detection paradigms, incorporating a denoising interaction unit to jointly optimize classification and localization. As a result, it reduces missed detections and false positives in densely occluded underwater scenes, improving detection accuracy beyond conventional heads in complex environments.
- A multimodal underwater object detection dataset, containing synchronized RGB and depth images, is constructed. The effectiveness of the proposed modules and the overall model is validated through comparative and ablation experiments. The results show that UW-DualDet achieves a mean Average Precision (mAP) of 86.7% and demonstrates strong detection performance and robustness in three representative underwater scenarios: bright, dim, and densely populated scenes. These findings highlight its practical potential for applications such as underwater intelligent operations and marine ecological monitoring.

The rest of this paper is structured as follows. The Section 2 reviews studies on underwater object detection as well as relevant work on multimodal feature enhancement and fusion. In section 3, the relevant theories of UDFF, UFEM and D26Head methods are introduced. The overall structure, which describes UW-DualDet. Section 4 presents the experiments conducted with related methods on the datasets. Finally, Section 5 concludes the paper.

2 Related works

2.1 Underwater object detection

Object detection methods for complex underwater scenes can be broadly categorized into three approaches: lightweight methods based on one-stage networks, high-accuracy methods using two-stage networks, and feature enhancement methods utilizing Transformer architectures Gao et al. (2024); Song et al. (2023); Guo et al. (2024).

One-stage detection algorithms, exemplified by the YOLO series, have become the dominant architecture for underwater engineering applications due to their favorable balance between accuracy and speed. Existing research primarily focuses on two areas: customized optimization for aquaculture scenarios and lightweight design with small-object enhancement for general underwater scenes. Yu et al. (2025) proposed TMAE-YOLO, which achieves cross-frame feature compensation and multi-scale feature optimization via a temporal multi-frame aggregation module and a top-down progressive fusion feature pyramid. To address small-object feature loss and severe sample imbalance in turbid water, Tu et al. (2025) proposed YOLOv10-UDFishNet. This method significantly improves detection accuracy and recall of diseased fish through comprehensive pipeline optimization, including auxiliary-branch backbone enhancement, highlevel feature-guided adaptive fusion, and loss function refinement. Wei et al. (2026) developed YOLO-LD to meet the real-time monitoring needs for dead fish in recirculating aquaculture systems with mixed cultures. This model balances speed and accuracy by reconstructing the backbone network and introducing a global-local interactive feature fusion module along with a lightweight detection head. Zhou et al. (2025) introduced UW-YOLO-Bio, which balances detection accuracy and speed via a global context perception module, efficient downsampling, and a regional context feature pyramid. Ye et al. (2025) proposed YOLOv11-MSE, integrating a multi-scale dilated attention module with a lightweight Slim-Neck architecture to handle dense underwater small targets and severe category imbalance. Sun et al. (2025) introduced AGS-YOLO, which significantly improves detection performance under limited computational resources by optimizing multi-scale attention, lightweight convolution, and a cross-scale feature enhancement network. Yin et al. (2026) proposed TFDF-YOLO to address target edge degradation and localization bias through a spatial-frequency decoupling module, multi-scale fusion strategy, and optimized anchor-box regression loss. This design satisfies the need for precise positioning in underwater wireless power transfer docking for unmanned underwater vehicles (UUVs). Overall, YOLO-based one-stage detection methods effectively balance detection accuracy and real-time performance in underwater scenes and have become the dominant solution for underwater engineering applications. However, these methods rely solely on single-modal optical images, which degrade severely in underwater environments characterized by extreme turbidity, low illumination, and strong scattering. Consequently, their detection performance is compromised, limiting adaptability to complex and dynamic real-world underwater scenarios.

Two-stage detection algorithms, represented by the R-CNN series, offer higher accuracy and robustness in complex underwater scenes due to their refined feature processing and object discrimination capabilities Song et al. (2023). As a result, they have become the dominant architecture for high-precision underwater detection tasks. Song et al. (2023) proposed Boosting R-CNN, which

integrates a RetinaRPN region proposal network, a probabilistic inference pipeline, and a reweighted hard-example mining strategy to enable deep fusion of two-stage detection information and effective error correction. Zhao et al. (2021) developed Composited FishNet, a composite detection framework based on Cascade R-CNN, to address fish detection and species recognition in low-quality underwater videos. This framework suppresses underwater background interference using a composite backbone network, optimizes multi-scale feature fusion with an enhanced path aggregation network, and alleviates sample imbalance through a composite loss function. Two-stage methods, with more refined detection processes, substantially improve detection accuracy in complex underwater environments. However, their inherent two-stage computational framework results in substantially higher computational overhead, making it difficult to meet real-time operational requirements for underwater robots. Additionally, these methods still rely on single-modal optical images, which cannot guarantee stable and robust detection in extreme underwater conditions.

Transformer architectures, with their ability to model global context and capture long-range dependencies, offer a promising avenue for feature enhancement in underwater object detection and have become a hot research topic in recent years. To address the detection of arbitrarily oriented targets in complex underwater environments, Gao et al. (2024) proposed PE-Transformer. This method uses CSWin-Transformer as the backbone, strengthens bidirectional fusion of high- and low-level features through a path enhancement module, employs Cross-PAFPN to optimize multi-scale feature interaction, and introduces an adaptive point-representation detection head for precise target detection. To tackle few-shot detection in underwater sonar images with sample scarcity, category imbalance, and the presence of weak, small, and blurred targets, Yang et al. (2026) proposed FS2-DETR. This method enhances feature representation and category discrimination for weak and small targets through a decoder-guided feature enhancement and compensation mechanism, dual-level visual prompt enhancement, and a multi-stage progressive training strategy. Transformer architectures significantly improve global feature modeling in underwater detection and show clear advantages in few-shot learning and target occlusion scenarios. However, these methods are primarily optimized for single-modal data, such as optical or sonar images alone, and cannot fully resolve imaging failure in extreme underwater environments through multimodal complementarity. Research on multimodal information complementarity is crucial for improving underwater object detection performance.

2.2 Underwater images feature fusion

Recent studies have increasingly explored feature fusion strategies to enhance the representational capacity of underwater object detection networks. Chen et al. (2026) proposed ITT-YOLO, which integrates a triple-branch parallel attention module with a triple feature encoding module for multiscale feature aggregation. The attention module combines global self-attention, multi-scale local attention, and identity-preserving pathways, thereby improving the network's ability to capture complementary contextual information. Li et al. (2025b) developed SGL-YOLO by introducing a global-local cross-layer aligned BiFPN. This structure incorporates both global and local spatial attention branches, enabling the model to capture long-range dependencies and fine-grained spatial details during cross-scale feature fusion. Lv and Pan (2025) proposed LUOD-YOLO, which employs a dynamic local-global feature fusion attention mechanism based on adaptive pooling. In this method, local and global attention maps are integrated through dynamically learned weights, while bidirectional compensation paths further enhance cross-scale feature interaction. Liang et al. (2025) proposed RCF-YOLO, in which positional information is embedded into channel features through a coordinate enhancement attention module. The model further improves feature extraction using receptive field attention convolution, which adaptively adjusts convolutional kernel weights according to the spatial characteristics of different receptive fields. Li et al. (2025a) introduced UAED-Net, which establishes deep interactions between image enhancement and object detection tasks through a task-adaptive feature interactor that dynamically generates fusion weight maps. In addition, its multi-dimensional cross-scale fusion module encodes and integrates features along the channel, width, and height dimensions. Jin et al. (2026) proposed HATBformer, which, although primarily designed for underwater image enhancement, demonstrates effective feature integration through a detail enhancement branch that fuses coordinate-aware features to restore degraded image details.

Although these studies have advanced attention-based, cross-scale, and task-interactive feature fusion, they remain limited to single-modal RGB inputs. Under severe turbidity or low-illumination conditions, RGB textures may be substantially degraded, and fusion strategies operating solely within the RGB modality cannot recover information that is absent from the original input. In contrast, depth images can provide complementary geometric and structural cues that are less dependent on visual texture, yet such information remains insufficiently explored in existing underwater detection frameworks. To address this limitation, we propose UW-DualDet, a dual-modal underwater detection framework that integrates RGB texture information with depth-based geometric cues. Specifically, the proposed Underwater Denoise Feature Fusion (UDFF) module adaptively fuses RGB and depth features while explicitly suppressing noise interference. Furthermore, the Underwater Feature Enhancement Module (UFEM) strengthens the fused representations through multi-scale convolution and multi-dimensional weighting. These two modules jointly alleviate the inherent limitations of single-modal RGB fusion and improve the robustness of underwater object detection under degraded visual conditions.

2.3 Multimodal feature enhancement and fusion

In complex underwater environments characterized by low illumination, turbidity, and strong scattering, single-modality RGB imaging is highly susceptible to texture degradation, severe contrast loss, and target information loss. These factors substantially degrade detection performance and limit the achievable

accuracy of models relying solely on single-modal data Xu et al. (2025).

In contrast, multimodal data, such as RGB-infrared and RGB-depth, provide complementary information on appearance, spatial structure, and environmental contours, compensating for the limitations of a single modality at the data level Yang et al. (2025). As a result, multimodal feature enhancement and fusion have become a central research focus for overcoming performance bottlenecks in complex scenes and improving model robustness. The primary objective is to enable efficient fusion of complementary cross-modal information, while suppressing redundant inter-modal interference and enhancing discriminative target features, ultimately improving detection accuracy.

To address the low efficiency of multimodal feature fusion, Li et al. (2023) and Cao et al. (2023) proposed lightweight cross-modal fusion modules, both named CSSA. These modules were designed around coordinated optimization in both channel and spatial domains. Specifically, an ECAbased channel switching mechanism enables efficient cross-channel interaction between RGB and infrared features, while a parameter-free spatial attention module enhances foreground targets and suppresses redundant background noise. With minimal computational overhead, these methods achieve efficient multimodal fusion and feature enhancement, improving both detection accuracy and inference efficiency.

To tackle inter-modal interference and feature loss during cross-modal fusion, Chen et al. (2024b) proposed RI-YOLO, an end-to-end RGB-infrared multimodal detection framework. The framework constructs a full-chain feature fusion and enhancement system. First, the G&M-DI module performs twostage fusion of global and interdependent characteristics across the two modalities, enhancing intra-modal features. Then, the AWI module adaptively fuses deep multi-scale features, reducing redundant crossmodal interference and exploiting bimodal complementarity more effectively. These components together enable full-chain target feature enhancement and provide an efficient strategy for optimizing end-to-end multimodal detection frameworks.

To address the challenges in complex underwater scenes, including the limited representational capacity of a single RGB modality, difficulties in modeling both pixel-level details and high-level semantic information, and severe underwater background interference, Yang et al. (2025) proposed UMIS-YOLO, a multimodal instance segmentation framework based on the YOLO architecture. The framework uses a dualbackbone hierarchical multi-scale cross-modal fusion system. Two parallel backbones extract texture-color features from RGB images and boundary-structural features from depth images, respectively. The FDFEF module maps bimodal features into the frequency domain, where adaptive cross-modal fusion of amplitude and phase components occurs after modality-specific denoising and enhancement. Additionally, the RFF module achieves cross-level fusion between shallow bimodal features and deep semantic features through residual connections. By combining frequency-domain adaptive enhancement with spatial-domain residual detail enhancement, the method strengthens multidimensional target features while suppressing underwater background noise. As a result, it achieves deep complementary fusion and discriminative enhancement of RGB-depth features, significantly improving accuracy and robustness in instance segmentation in complex underwater scenes.

Overall, existing studies have made substantial progress in multimodal feature enhancement and fusion. However, for underwater object detection tasks with unique scene characteristics, existing methods still exhibit three major limitations. First, most fusion schemes are primarily designed for RGB-infrared modalities and show limited adaptability to the heterogeneous characteristics of underwater RGB-depth data. Consequently, they cannot effectively address modality-specific issues such as texture degradation in RGB images and edge noise in depth maps. Third, existing underwater multimodal fusion methods mainly focus on pixel-level optimization for instance segmentation, lacking dedicated solutions for underwater object detection, particularly in scenarios with a high proportion of small targets and severe feature degradation.

To address these limitations, this paper proposes UW-DualDet, an underwater RGB-depth bimodal detection network specifically designed for the key challenges of underwater scenes. First, the underwater feature enhancement module, UFEM, is introduced to strengthen the representation of low-contrast and small underwater targets while suppressing background noise caused by water scattering. Second, the underwater denoising feature fusion module, UDFF, is designed to achieve adaptive complementary fusion of RGB texture and depth structural information, effectively reducing redundant cross-modal interference. Overall, the proposed method substantially improves detection accuracy and robustness in extreme underwater scenes, providing an effective solution for multimodal object detection in complex underwater environments.

3 Materials and methods

In this section, a network overview is provided in Section 3.1, following which the proposed Underwater Denoise Feature Fusion (UDFF) is described in Section 3.2. The proposed Underwater Feature Enhancement Module(UFEM) is explained in Section 3.3. The proposed Denoise YOLO26 Detection Head (D26Head) is explained in Section 3.4.

3.1 Underwater dual-modal detection network

Object detection in underwater environments faces significant challenges: light absorption and scattering lead to color distortion, reduced contrast, and blurred details in RGB images, while depth images in turbid water are prone to noise and missing regions, limiting the effectiveness of single-modality scene understanding.

To address these issues, this study proposes an underwater RGB-D multimodal object detection model. Built upon the end-to-end detection framework, the proposed model incorporates an Underwater Denoise Feature Fusion module (UDFF), a D26Head, and an Underwater Feature Enhancement Module (UFEM). A dual-layer fusion strategy is adopted at P_4 and P_5 to enable effective interaction between RGB and depth features. Furthermore, the denoising detection head D26Head supports end-to-end inference

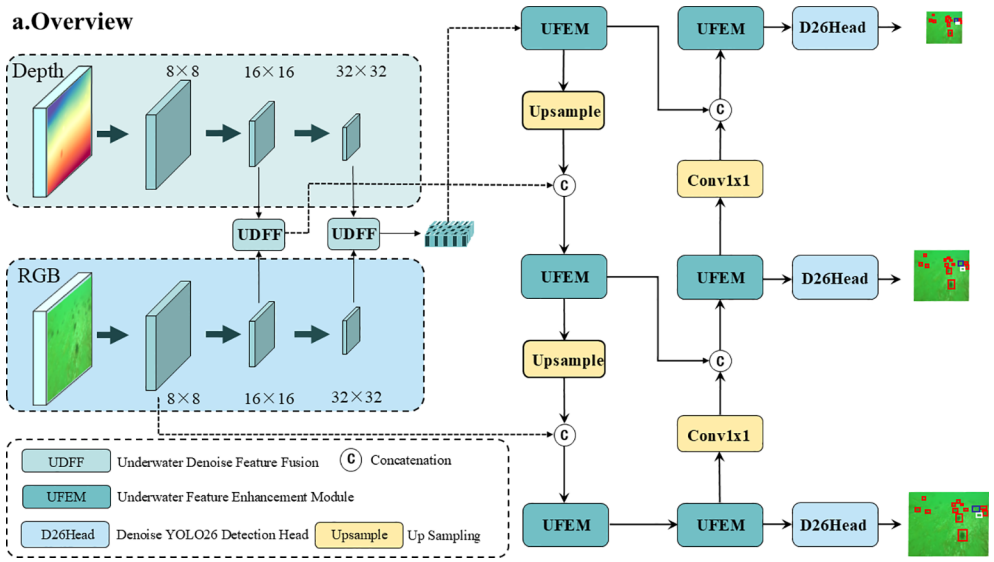

FIGURE 2
The structure of underwater dual-modal detection network.

without Non-Maximum Suppression (NMS). The overall architecture is illustrated in Figure 2.

Given an input RGB image $I_{RGB} \in \mathbb{R}^{3\times H\times W}$, the monocular depth estimation model DepthAnythingV2-Large generates a single-channel relative depth map $I_{Depth\_raw} \in \mathbb{R}^{1\times H\times W}$. To match the input dimensionality of the dual-parallel backbone network, the single-channel depth map is replicated along the channel dimension to obtain a three-channel depth input $I_{Depth} \in \mathbb{R}^{3\times H\times W}$. The RGB and depth backbones share the same convolutional downsampling architecture, comprising five hierarchical levels from P_1 to P_5. The feature map at the l-th level is defined in Equation (1):

$$\mathcal{F}^{(l)}_{\mathrm{RGB}} = \Phi^{(l)}_{\mathrm{RGB}}(I_{\mathrm{RGB}}), \quad \mathcal{F}^{(l)}_{\mathrm{Depth}} = \Phi^{(l)}_{\mathrm{Depth}}(I_{\mathrm{Depth}}), l \in \{3,4,5\} \quad (1)$$

Where, $\Phi^{(l)}$ denotes the feature extraction function at the l-th stage of the YOLO26 backbone, which comprises convolutional layers, C3k2 modules, and a Spatial Pyramid Pooling Fast (SPPF) structure. The SPPF module incorporates residual connections to enhance gradient flow. Due to the presence of noise in underwater depth maps, this study introduces an Underwater Denoising Feature Fusion (UDFF) module at the P_4 and P_5 levels. This module performs cross-modal fusion of RGB and depth features while suppressing depth noise. Let $\mathcal{F}\mathrm{RGB}^{(l)}$ and $\mathcal{F}\mathrm{Depth}^{(l)}$ denote the input features. The UDFF module first applies Gaussian filtering and adaptive thresholding to denoise the depth features, and then performs feature interaction via cross-modal channel attention, producing the fused feature. These processes can be computed in Equation (2):

$$\mathcal{F}^{(l)}_{\mathrm{fuse}} = \mathrm{UDFF}\left(\mathcal{F}^{(l)}_{\mathrm{RGB}}, \mathcal{F}^{(l)}_{\mathrm{Depth}}\right), \quad l \in \{4,5\} \quad (2)$$

The internal mechanism of the UDFF module can be described as follows. First, a denoising operator D is applied to the depth features, i.e., $\tilde{\mathcal{F}}_{\mathrm{Depth}} = \mathcal{D}(\mathcal{F}_{\mathrm{Depth}})$, where $\mathcal{D}$ consists of Gaussian filtering and adaptive thresholding. Subsequently, the RGB features and the denoised depth features are projected into a shared channel dimension C, and a cross-modal attention matrix $\mathbf{A} \in \mathbb{R}^{C\times C}$ is computed. The fused features are then obtained via residual connections, which can be mathematically expressed in Equation (3):

$$\mathbf{A} = \mathrm{Softmax}\left(\frac{Q}{\|Q\|} \cdot \frac{K}{\|K\|}^{\top}\right), \qquad \mathcal{F}^{(l)}_{\mathrm{fuse}} = \mathrm{Proj}_{\mathrm{out}}(\gamma \cdot AV + \mathcal{F}_{RGB}) \quad (3)$$

Where, Q, K, and V denote the query, key, and value tensors obtained from the projected RGB and depth features, respectively; γ is a learnable fusion parameter, and $\mathrm{Proj}_{\mathrm{out}}$ aligns the output channels with the RGB branch. After UDFF-based fusion, the P_4 and P_5 levels yield the fused features $\mathcal{F}^{(4)}_{\mathrm{fuse}}$ and $\mathcal{F}^{(5)}_{\mathrm{fuse}}$, which are further processed through the C3k2 and C2PSA modules to extract semantic information, resulting in the enhanced fused features $\mathcal{F}_{P_4}$ and $\mathcal{F}_{P_5}$.

To fully leverage multi-scale information, the model employs a bidirectional feature fusion neck composed of a Feature Pyramid Network (FPN) and a Path Aggregation Network (PAN), wherein an Underwater Feature Enhancement Module (UFEM) replaces conventional convolutions for feature refinement. Let the input feature be $\mathcal{X} \in \mathbb{R}^{C\times H\times W}$. UFEM first captures local details via depthwise separable convolutions and then incorporates an Efficient Multi-scale Attention (EMA) mechanism to enhance salient regions in underwater scenes, producing the enhanced feature. The process can be computed in Equation (4):

$$\mathcal{X}_{\mathrm{enhanced}} = \mathrm{UFEM}(\mathcal{X}) = \mathrm{EMA}(\mathrm{DWConv}(\mathcal{X})) + \mathcal{X} \quad (4)$$

Where, DWConv denotes depthwise separable convolution, and EMA refers to the Efficient Multi-scale Attention mechanism, which computes channel attention and multi-scale spatial attention in parallel, adaptively weighting the fusion of global contextual information and local details. Within the neck network, the FPN

path performs progressive upsampling from $\mathcal{F}_{P5}$ and fuses features with $\mathcal{F}_{P4}$ and RGB_{P3}, while the PAN path aggregates features from all levels in a bottom-up manner. UFEM refines the features at each fusion node, resulting in enhanced multi-scale feature maps: $\mathcal{Y}_{P3}$, $\mathcal{Y}_{P4}$, and $\mathcal{Y}_{P5}$.

Finally, the three-scale feature maps are fed into D26Head, which adopts a dual-branch mechanism enabling end-to-end inference. Let the input features be $\mathcal{Y} \in \mathcal{Y}_{P3}, \mathcal{Y}_{P4}, \mathcal{Y}_{P5}$. The detection head consists of a one-to-many branch for training and a one-to-one branch for inference. During training, both branches compute losses in parallel. The one-to-one branch utilizes Hungarian matching to enforce strict one-to-one label assignment, guiding the network to learn de-duplication capability. During inference, only the one-toone branch is retained, producing a fixed number of candidate boxes $\in \mathbb{R}^{N\times 6}$, where $N = 300$. Each candidate box contains bounding box coordinates, object confidence, and class probabilities. The prediction process can be defined as Equation (5):

$$\mathcal{P} = \text{D26Head}_{\text{infer}}(\mathcal{Y}_{P3}, \mathcal{Y}_{P4}, \mathcal{Y}_{P5}) \tag{5}$$

This architecture suppresses underwater depth noise via UDFF and enhances salient feature regions in underwater scenes through UFEM. Combined with the end-to-end detection capability of YOLO26 Sapkota et al. (2025), it achieves efficient and robust multi-modal object detection in complex underwater environments.

3.2 Underwater denoise feature fusion

The Underwater Denoising Feature Fusion (UDFF) module is a core component specifically designed for underwater RGB-D multi-modal feature fusion. Built upon the classical cross-modal attention mechanism, it is optimized to address the degradation and noise characteristics commonly observed in underwater depth images. In underwater environments, depth images are often affected by scattering, absorption, suspended particles, and inaccurate depth estimation, leading to non-Gaussian noise, missing values, blurred structures, and outlier regions. Directly using such degraded depth images in feature fusion may severely interfere with cross-modal information interaction and weaken the reliability of RGB-D representation learning. To mitigate this issue, UDFF introduces lightweight denoising preprocessing, numerical stability enhancements, and optional local spatial attention, thereby refining underwater depth features and enabling more robust cross-modal fusion.

As shown in Figure 3, let the inputs be the RGB feature $X_{RGB} \in \mathbb{R}^{B\times C_{RGB}\times H\times W}$ and the depth feature $X_{Depth} \in \mathbb{R}^{B\times C_{Depth}\times HmesW}$. UDFF first applies denoising preprocessing to the depth features. The denoising operator $\mathcal{D}$ combines Gaussian filtering with adaptive thresholding: the depth features are first smoothed using a Gaussian filter to suppress high-frequency noise. Next, within a local 3 × 3 neighborhood, the mean μ and standard deviation σ_{local} are computed. An adaptive threshold is set as Equation (6):

$$\tau = \mu + 2\sigma_{local} \tag{6}$$

Outlier points exceeding this threshold are replaced with the local mean. This process effectively suppresses outlier noise in underwater depth maps while preserving edge structures. The process can be defined in Equation (7):

$$\tilde{X}_{Depth} = \mathcal{D}(X_{Depth}) \tag{7}$$

Subsequently, the RGB features and the denoised depth features are projected into a shared channel space C_{common} via 1 × 1 convolutions, yielding the projected features F_{RGB} and F_{Depth}.

The cross-modal attention mechanism captures the correlations between RGB and depth features along the channel dimension. First, the features are reshaped into sequential form to construct the query Q, key K, and value V tensors. To prevent numerical instability, both the query and key are L2-normalized. Subsequently, the channel attention matrix is computed and normalized by applying the Softmax function along the channel dimension. The process can be computed in Equation (8):

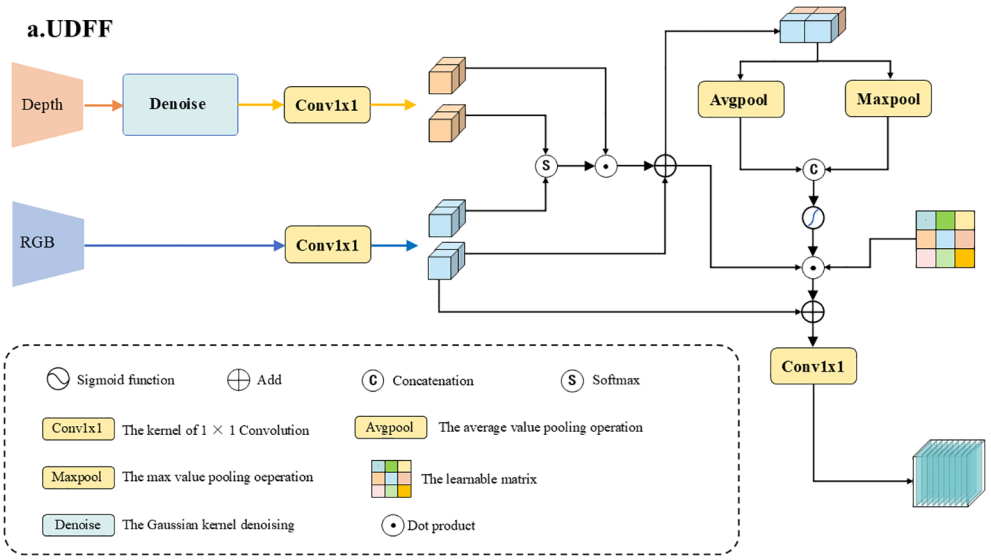

FIGURE 3
The structure of underwater denoise feature fusion (UDFF).

$$A = Softmax_{dim=-1}\left(clamp\left(\hat{Q} \cdot \hat{K}^{\top}, 0, 1\right)\right) \tag{8}$$

where $\hat{Q}$ and $\hat{K}$ denote the L_2-normalized query and key, respectively. The clamp operation truncates negative correlation values to zero and is designed to suppress noise-induced conflicting signals between the RGB and depth modalities in underwater environments. As discussed below, negative correlations in underwater cross-modal data are more likely to arise from noise and artifacts than from genuine complementary information, and removing them improves fusion robustness. The attention information is obtained via A · V, which is then reshaped back to the spatial dimensions and fused with the original projected RGB features through a residual connection. The processes can be defined as Equation (9):

$$\mathcal{F}_{\text{fused}} = \gamma \cdot \text{Attn}(\mathcal{F}_{\text{RGB}}, \mathcal{F}_{\text{Depth}}) + \mathcal{F}_{\text{RGB}} \tag{9}$$

where, γ is a learnable fusion parameter, initialized to 0, which ensures that the original feature distribution is preserved during the early stages of training.

To enhance the perception of local salient regions in underwater scenes, UDFF incorporates a lightweight spatial attention mechanism. This branch computes the mean and maximum values along the channel dimension, concatenates them, and passes the result through a 3 × 3 convolution followed by a Sigmoid activation to generate a spatial weight map. The map is subsequently used to reweight and enhance the fused feature. The process can be formulated as Equation (10):

$$\mathcal{F}_{\text{enhanced}} = \mathcal{F}_{\text{fused}} + \gamma_{\text{spatial}} \cdot (\mathcal{F}_{\text{fused}} \odot \text{Sigmoid}(\text{Conv}_{3\times3}(\text{Cat}(\text{Mean}(\mathcal{F}_{\text{fused}}), \text{Max}(\mathcal{F}_{\text{fused}})))))) \tag{10}$$

where, γ_{spatial} is a learnable spatial attention weight, $\odot$ denotes element-wise multiplication, and Cat represents channel-wise concatenation. This design effectively emphasizes critical regions of underwater objects, thereby enhancing the discriminative capability of the features.

Finally, the enhanced feature is projected to the RGB branch channel dimension via the output projection Proj_{out}, producing the final fused feature. The process can be defined in Equation (11):

$$Y = \text{Proj}_{\text{out}}(\mathcal{F}_{\text{enhanced}}) \tag{11}$$

Through this design, the UDFF preserves the lightweight and efficient properties of cross-modal attention while incorporating depth feature denoising and spatial attention enhancement. It specifically addresses the degradation of underwater depth features and improves the model's perception of salient underwater targets, providing high-quality fused feature representations for subsequent multi-scale feature fusion and object detection.

3.3 Underwater feature enhancement module

The Underwater Feature Enhancement Module (UFEM) is an efficient module specifically designed to address feature degradation in underwater environments. Underwater images often suffer from contrast reduction and blurred details due to scattering and absorption, hindering traditional feature extraction methods from effectively capturing salient regions. As illustrated in Figure 4, UFEM incorporates an Efficient Multi-scale Attention (EMA) mechanism to perform multi-scale contextual modeling and joint channel–spatial enhancement of underwater features, thereby improving the model's perception of underwater targets.

Let the input feature be $X \in \mathbb{R}^{B\times C_{\text{in}}\times H\times W}$ UFEM first expands the channel dimension to $2C$ via a 1 × 1 convolution, where $C = C_{\text{in}} \cdot e$, and e denotes the expansion factor (default: 0.5). The process can be defined as Equation (12):

$$X_{\text{exp}} = \text{Conv}_{1\times1}(X) \in \mathbb{R}^{B\times2C\times H\times W} \tag{12}$$

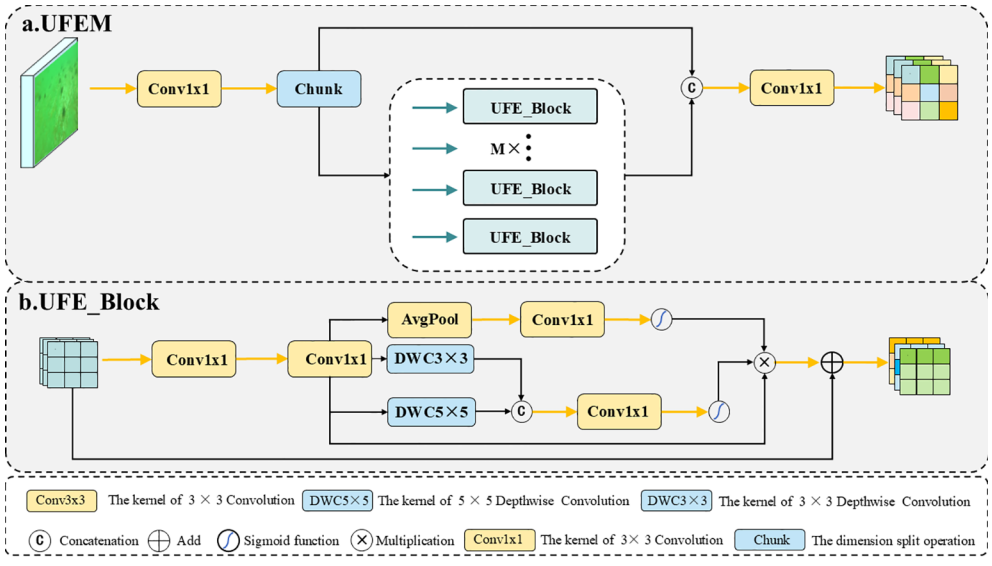

FIGURE 4
The structure of underwater feature enhancement module (UFEM).

Subsequently, the expanded feature is split along the channel dimension into two parts, X_0 and X_1, each containing C channels. The branch $\mathbf{X}_1$ is then passed through n UFE_Bblock modules for deep feature extraction. The process can be computed as Equation (13):

$$X_1^{(k)} = \text{UFE\_Block}(X_1^{(k-1)}), \quad k = 1, 2, \ldots, n \tag{13}$$

where, $X_1^{(0)} = X_1$. All intermediate features are concatenated with the initially split features along the channel dimension. The process can be defined in Equation (14):

$$X_{\text{cat}} = \text{Concat}(X_0, X_1^{(0)}, X_1^{(1)}, \ldots, X_1^{(n-1)}) \in \mathbb{R}^{B\times(2+n)C\times H\times W} \tag{14}$$

Finally, the channel dimension is projected back to the target dimension C_{out} via a 1 × 1 convolution, yielding the enhanced feature. The process can be computed as Equation (15):

$$Y = \text{Conv}_{1\times1}(X_{\text{cat}}) \in \mathbb{R}^{B\times C_{\text{out}}\times H\times W} \tag{15}$$

The *UFE_Block* serves as the core building unit of the UFEM module. Its design follows the C3k2 architecture, enabling effective extraction of underwater features through multi-branch feature reuse and progressive feature enhancement. Within the *UFE_Block*, the input f first undergoes two stages of channel-wise semantic filtering to obtain f_1, which can be formulated as Equation (16):

$$f_1 = Conv_{1\times1}(Conv_{1\times1}(f)) \tag{16}$$

The feature f_1 is processed through parallel channel-wise weighting and multi-scale weighting, enabling multi-dimensional semantic enhancement of the feature. These processes can be expressed as Equations 17–19:

$$W_1 = Sigmoid(Conv_{1\times1}(AvgPool(f_1))) \in \mathbb{R}^{B\times C\times1\times1} \tag{17}$$

$$W_2 = Sigmoid(Conv_{1\times1}(Concat(DWC_{3\times3}(f_1), DWC_{5\times5}(f_1)))) \in \mathbb{R}^{B\times1\times H\times W} \tag{18}$$

$$f_{enhancement} = f_1 * W_1 * W_2 \tag{19}$$

where, the weight $w_1 \in \mathbb{R}^{B\times C\times1\times1}$ is the channel attention weight, and the weight $w_2 \in \mathbb{R}^{B\times1\times H\times W}$ is the spatial attention weight. $f_{enhancement} \in \mathbb{R}^{B\times C\times H\times W}$. The weight-enhanced underwater features are then added to the original features via a residual connection, preserving the original foreground semantic information in underwater scenes. The process can be computed as Equation (20):

$$f_{out} = f_{enhancement} + f \in \mathbb{R}^{B\times C\times H\times W} \tag{20}$$

Through this design, the UFEM module mitigates feature degradation in underwater object detection and provides robust feature representations for accurate target localization and classification. By capturing multi-scale contextual information through depthwise separable convolutions and progressive feature reuse, UFEM enhances the model's ability to perceive underwater targets of varying sizes, particularly small targets that are easily obscured by background noise. The integrated EMA mechanism jointly models channel-wise and spatial feature dependencies, adaptively enhancing the saliency of low-contrast targets while suppressing background interference caused by water scattering and suspended particles. This design improves the discriminative capacity of features for blurred and partially visible underwater targets. It also facilitates cross-scale information interaction in the subsequent feature pyramid network and supports accurate bounding-box regression and classification in the detection head.

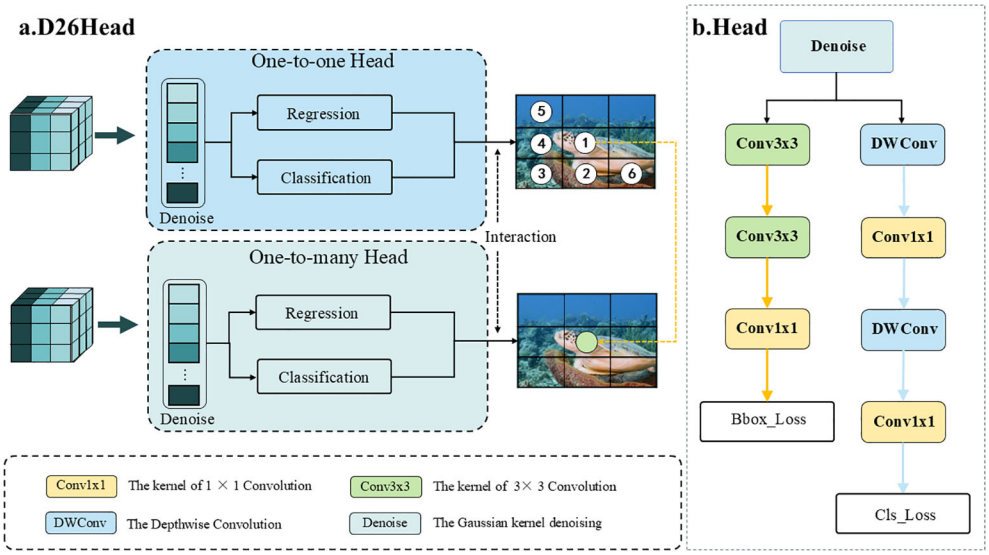

FIGURE 5
The structure of denoise YOLO26 detection head (D26Head).

3.4 Denoise YOLO26 detection head

The D26Head detection head is developed based on the end-to-end detection architecture of YOLO26 Sapkota et al. (2025). While retaining the dual-branch training mechanism and the simplified Distribution Focal Loss, it incorporates a Gaussian denoising preprocessing step prior to feature input to enhance robustness against noisy features in underwater environments.

The D26Head architecture is shown in Figure 5. Let the input to the detection head be a set of multiscale feature maps $\mathcal{Y} = Y_{P3}, Y_{P4}, Y_{P5}$, where $Y_{P3} \in \mathbb{R}^{B\times C_3\times H_3\times W_3}$, $Y_{P4} \in \mathbb{R}^{B\times C_4\times H_4\times W_4}$, and $Y_{P5} \in \mathbb{R}^{B\times C_5\times H_5\times W_5}$ correspond to feature maps at small, medium, and large scales, respectively. To suppress noise introduced by underwater environments, D26Head first applies Gaussian denoising preprocessing to each scale feature map. For an arbitrary scale feature $Y \in \mathbb{R}^{B\times C\times H\times W}$, the denoising operation is implemented via convolution with a Gaussian kernel, producing the smoothed feature $\tilde{Y}$ The process is defined in Equation (21):

$$\tilde{Y} = Y \star G_\sigma, \quad G_\sigma(u,v) = \frac{1}{2\pi\sigma^2} e^{-\frac{u^2+v^2}{2\sigma^2}} \tag{21}$$

where, G_σ denotes the Gaussian kernel, σ is a configurable parameter, and $*$ represents the convolution operation. This operation effectively suppresses high-frequency noise in the feature maps while preserving structural information, thereby providing cleaner feature inputs for the subsequent detection head.

After denoising, the feature maps $\tilde{Y}$ are fed into two parallel branches: a One-to-many Head and a One-toone Head. Both heads share an identical internal structure, each consisting of a feature transformation stage followed by parallel classification and regression sub-branches. The denoised feature $\tilde{Y}$ is first processed through two consecutive 3 × 3 convolutional layers, each comprising a convolution operation followed by Batch Normalization and SiLU activation. This transformation is expressed as Equation (22):

$$P_{reg} = Conv_{1\times1}(SiLu(BN(Conv_{3\times3}(SiLU(BN(Conv_{3\times3}(\tilde{Y})))))))) \tag{22}$$

where $Conv_{3\times3}$, $Conv_{1\times1}$ denote the 3 × 3 and 1 × 1 convolution operation. BN represents Batch Normalization, and $SiLU(x) = x \cdot \sigma(x)$ is the sigmoid linear unit activation function. The transformed feature $P_{reg} \in \mathbb{R}^{B\times4\times H\times W}$ represents the final output of the regression branch, which encodes the spatial prediction information of the underwater objects and the four output channels represent the predicted bounding box coordinates (b_x, b_y, b_w, b_h).

The classification sub-branch processes $\tilde{Y}$ through a 3 × 3 depthwise separable convolution followed by a 1 × 1 convolution to predict class probabilities at each spatial location. The depthwise separable convolution consists of a depthwise convolution and a pointwise convolution, which can be expressed as Equation (23):

$$F_{mid} = Conv_{1\times1}(DWC_{3\times3}(\tilde{Y})) \tag{23}$$

where $DWC_{3\times3}$ denote the 3 × 3 depthwise convolution. The final class prediction is processed as Equation (24):

$$P_{cls} = Conv_{1\times1}(DWC_{3\times3}(F_{mid})) \in \mathbb{R}^{B\times N_{cls}\times H\times W} \tag{24}$$

where N_{cls} denotes the number of underwater object classes.

The one-to-many branch and the one-to-one branch share identical structures but differ in their label assignment strategies. The one-to-many branch adopts a conventional label assignment strategy, where each ground-truth object is matched with multiple positive samples, thereby providing abundant supervisory signals for network training. In contrast, the one-to-one branch employs the Hungarian matching algorithm to achieve strict one-to-one label assignment, in which each ground-truth object is matched with only the best prediction, thereby enforcing the network to learn de-duplication capability. During training, for the one-to-many branch, the loss function $\mathcal{L}_{one2many}$ is formulated as a weighted combination of classification loss and regression loss. Similarly, for the one-to-one branch, the loss function $\mathcal{L}_{one2one}$ also consists of classification and regression terms, but the cost matrix is computed using Hungarian matching. The overall training loss is computed in Equation (25):

$$\mathcal{L}_{total} = \mathcal{L}_{one2many} + \mathcal{L}_{one2one} \tag{25}$$

In the regression branch, D26Head adopts the simplified Distribution Focal Loss design of YOLO26 by setting $reg_{max} = 1$, directly predicting bounding box coordinates without requiring complex distribution modeling. For the predicted bounding box $b = (b_x, b_y, b_w, b_h)$ and the ground-truth box $\hat{b} = (\hat{b}_x, \hat{b}_y, \hat{b}_w, \hat{b}_h)$, the regression loss is defined as Equation (26):

$$\mathcal{L}_{reg} = 1 - \mathrm{GIoU}(b, \hat{b}) \tag{26}$$

The classification branch employs the Binary Cross-Entropy (BCE) loss. For the predicted class probability p and the ground-truth label $\hat{p}$, the classification loss is computed in Equation (27):

$$\mathcal{L}_{cls} = -\hat{p} \log(p) - (1-\hat{p}) \log(1-p) \tag{27}$$

During inference, D26Head retains only the one-to-one branch and directly outputs a fixed number of candidate boxes. For each scale of feature maps, the outputs of the regression and classification branches are decoded to obtain predicted bounding box coordinates and class confidences. Predictions from all scales are then merged, and the top N candidates with the highest confidence scores (N = 300) are selected as the final outputs, where each candidate comprises the bounding box coordinates, object confidence, and class label, represented as $\mathcal{P} \in \mathbb{R}^{N\times6}$. Since the one-to-one branch has learned deduplication capability during training via Hungarian matching, no Non-Maximum Suppression (NMS) is required during inference, enabling truly end-to-end detection and significantly reducing inference latency.

By incorporating Gaussian denoising preprocessing into the YOLO26 detection head, D26Head effectively enhances robustness against underwater feature noise. Combined with the dual-branch training mechanism and simplified regression design, it maintains efficient end-to-end inference while achieving precise underwater object detection. This design enables D26Head to better adapt to complex underwater environments, providing an efficient and reliable solution for underwater object detection tasks.

4 Experimental results

4.1 Multimodal underwater dataset

The upper bounds of accuracy, generalization capability, and robustness of underwater object detection algorithms are highly dependent on the modality completeness, scene coverage, category diversity, and annotation quality of the training dataset. Current mainstream public underwater object detection datasets face two major limitations: (1) most datasets provide only RGB images, lacking complementary modalities that enrich the available information, and (2) existing datasets predominantly focus on nearshore aquaculture species, with limited representation of coral reef ecosystems, underwater operations, and other scenarios. Additionally, these datasets lack sufficient samples under challenging conditions such as extreme turbidity, low illumination, and severe occlusion.

To address these limitations, this paper constructs a large-scale, multi-scenario, fully category-aligned RGB-Depth bimodal underwater object detection dataset. This dataset comprehensively covers key challenges in underwater object detection applications and provides high-quality data to support the training and performance evaluation of the proposed UW-DualDet bimodal detection network.

4.1.1 Data acquisition and scenario coverage

To ensure the realism of scenarios and the comprehensiveness of category coverage, this study acquires raw RGB images from two primary sources: *in-situ* underwater imaging and publicly available datasets. As a result, the constructed dataset encompasses three core underwater scenarios and 10 object categories.

The *in-situ* data acquisition focuses on nearshore aquaculture, a key application for underwater object detection, and is conducted in coastal aquaculture zones. As shown in Figure 6, the captured targets include four key economic marine species: echinus, holothurian, starfish, and scallop, which align with the core detection categories of the dataset. The acquisition system employs a professional underwater industrial camera equipped with a dedicated deep-water waterproof housing and an underwater white balance correction filter. Recording parameters are standardized to a resolution of 1920 × 1080 and a frame rate of 30 fps, with shooting depths ranging from 0.5 m to 15 m, covering typical operational depths in nearshore aquaculture. Raw images are extracted from recorded underwater videos through uniform-interval keyframe extraction. A strict sample screening process removes frames affected by motion blur, ghosting artifacts, and overexposure or underexposure. Ultimately, 5,000 high-quality *in-situ* images are retained, providing realistic and scenario-consistent nearshore aquaculture imagery.

To further enhance category coverage, scene diversity, and dataset scale, and to supplement categories not covered by *in-situ* acquisition, this study collects additional underwater image and video data from compliant public sources, including the Kaggle platform, the National Oceanic and Atmospheric Administration's (NOAA) marine imagery repository, and marine science outreach resources. This process supplements six additional object categories: fish, corals, diver, cuttlefish, turtle, and jellyfish, thus fully covering the 10 predefined detection categories of the dataset. The collected raw data undergo the same rigorous screening and preprocessing pipeline as the *in-situ* data, including deduplication, deblurring, removal of invalid samples, and lens distortion correction. A total of 7,000 compliant and valid images are retained, enriching rare category samples and non-aquaculture scenarios, such as coral reef ecosystems and underwater operations, as well as challenging conditions, including extreme turbidity, ultra-low illumination, and severe occlusion. The detection scenarios of the expanded underwater dataset are shown in Figure 7. This significantly enhances the dataset's scene diversity and category richness.

Through these acquisition and filtering processes, the dataset ultimately retains 12,000 valid RGB images, comprehensively covering three core application scenarios: nearshore aquaculture, coral reef ecological protection, and underwater engineering operations. This provides a solid foundation for subsequent bimodal data construction and annotation.

FIGURE 6
Four types of targets captured by the underwater camera: echinus, starfish, scallop, and holothurian.

FIGURE 7
The expanded underwater dataset includes four detection scenarios: dense-target, dim, turbid, and bright scenes, each accompanied by the corresponding Depth images aligned with the RGB images.

4.1.2 Specific underwater sub-scenario dataset

All scenario-specific evaluation subsets used in this study were constructed exclusively from the independent test set of our self-built underwater RGB-pseudo-depth dataset, without introducing any external datasets. The distribution of underwater targets across the three sub-scenario datasets is illustrated in Figure 8. To ensure independent and fair performance evaluation, all subsets were strictly mutually exclusive. That is, no image appeared in more than one subset, thereby reducing potential bias in cross-scenario comparisons.

The clear-water scenario subset consisted of underwater images captured under uniform natural illumination, with no visible turbidity, sharp target textures, and well-preserved edge details. This subset contained 820 test images and was used to establish the model's baseline detection performance under optimal underwater imaging conditions. The low-light deep-water scenario subset comprised images acquired under insufficient ambient illumination, characterized by low overall brightness, a low signal-tonoise ratio, and targets partially obscured by dark backgrounds. This subset included 560 test images and was designed to evaluate the model's robustness to severe low-light degradation, a common and critical challenge in deep-sea observation and nighttime underwater inspection. The dense-occlusion scenario subset contained underwater images with more than three overlapping targets, mutual occlusion among marine organisms, or targets partially obscured by reefs and other background structures. This subset included 650 test images and was used to assess the model's ability to detect partially occluded targets, which is essential for practical applications such as intensive aquaculture monitoring and biodiversity surveys in coral reef ecosystems.

4.1.3 RGB-depth bimodal data construction

To construct depth modality data that is spatially aligned with RGB images, this study explicitly employs the high-precision monocular depth estimation model DepthAnything-V2-Large Yang et al. (2024) to generate estimated relative pseudo-depth maps for all RGB images, rather than using depth data captured by physical underwater depth sensors. It is important to clarify that our dataset is therefore an RGB-pseudodepth bimodal dataset, not a hardware-synchronized RGB-D dataset. This approach was adopted out of necessity due to the extreme scarcity of large-scale, high-quality, temporally and spatially synchronized RGB-D datasets captured by physical underwater depth sensors, which remains a fundamental bottleneck limiting the development of multimodal underwater object detection research. More comprehensive discussions of the limitations are provided in Supplementary Section S2 of the Supplementary Materials.

All raw images are first uniformly resized to a resolution of 1080 × 1080 and undergo preprocessing steps, including underwater white balance correction, contrast enhancement, and de-scattering, to improve the stability of depth estimation. The preprocessed images are then fed into the depth model to generate single-channel depth maps, ultimately producing bimodal image pairs that are fully aligned with the RGB modality. As shown in Figure 7, the generated depth maps effectively capture spatial position, contour structure, and relative distance of targets, compensating for the limitations of

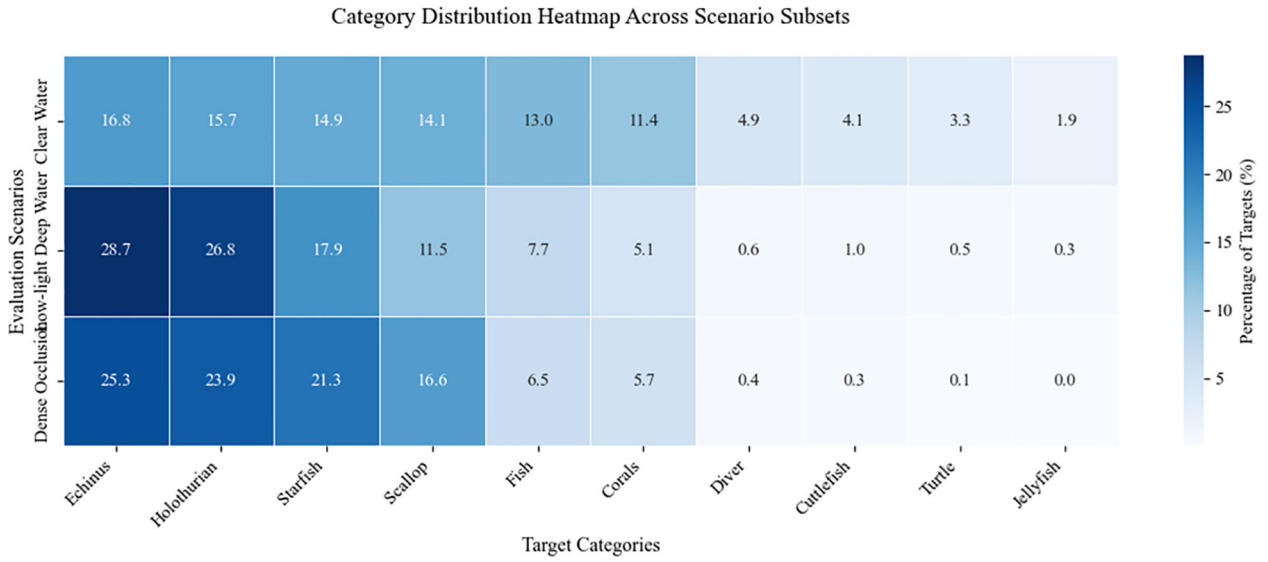

FIGURE 8
The heatmap of the distribution of underwater target counts across sub-scenario datasets.

underwater RGB images, such as texture degradation and target blurring under turbid and low-light conditions. This provides stable structural information for bimodal detection.

All depth map generation experiments are conducted on an NVIDIA RTX 4090–24 GB GPU using batch inference, ensuring both efficiency and consistency in data generation. The monocular depth estimation model DepthAnything-V2-Large generates a single-channel relative depth map $I_{Depth\_raw} \in \mathbb{R}^{1\times H\times W}$. To match the input dimensionality of the dual-parallel backbone network, the single-channel depth map is replicated along the channel dimension to obtain a three-channel depth input $I_{Depth} \in \mathbb{R}^{3\times H\times W}$. The model then uses two independent backbone networks to extract multi-scale features from the RGB image and the replicated three-channel depth map. The dataset partitioning strategy and the proportion of each object category can be found in Supplementary Section S1 of the Supplementary Materials.

4.2 Implementation details

All experiments in this study are conducted under a unified hardware and software environment to ensure fairness and reproducibility. The hardware platform includes an NVIDIA RTX 4090–24 GB GPU, a 16-core Intel Core i9 processor, and 64 GB of memory. The deep learning framework is built upon PyTorch 2.3 and CUDA 12.1, with Ubuntu 22.04 as the operating system.

The model training and evaluation configurations follow the official YOLOv11 settings Kotthapalli et al. (2025). The input image resolution is set to 640 × 640, with a training batch size of 16. The optimizer used is stochastic gradient descent (SGD) Tian et al. (2023), with an initial learning rate of 1×10^{-2}. A cosine annealing learning rate schedule is adopted, with a weight decay coefficient of 5×10^{-4} and a maximum of 300 training epochs.

Data preprocessing includes random horizontal flipping, random scaling, and Mosaic augmentation, aimed at improving the model's generalization capability in complex underwater environments.

4.3 Experimental evaluation metrics

This study constructs a comprehensive evaluation framework from three perspectives: detection accuracy, model complexity, and inference efficiency, to systematically assess the overall performance of the UWDualDet network. All metrics are defined according to commonly adopted standards in the object detection field, ensuring fairness and comparability of the experimental results.

For detection accuracy, mAP_{50} (mean Average Precision at an IoU threshold of 0.5) is used as the primary evaluation metric to measure overall detection performance. It is complemented by $mAP_{50:95}$, precision, and recall, which evaluate the model's multi-scale localization capability, false positive control, and missed detection suppression, respectively.

Regarding model complexity and real-time performance, the number of parameters (Params) and floatingpoint operations (GFLOPs) are used to quantify the model's lightweight characteristics and computational cost. Inference efficiency and hardware adaptability are evaluated using frames per second (FPS), per-frame inference latency, and peak GPU memory usage, with all metrics measured under a unified hardware and software environment.

4.4 Experiments on multimodal underwater dataset

4.4.1 Ablation study on the effectiveness of the bimodal architecture

To validate the effectiveness of the RGB-Depth bimodal fusion architecture and the rationale behind the dual-backbone parallel feature extraction design, this study compares detection performance under different modality inputs and baseline fusion architectures. The experimental results are presented in Table 1.

From the results, it is clear that the detection performance of the single RGB modality significantly outperforms that of the single

TABLE 1 Ablation analysis on the effectiveness of the bimodal architecture.

RGB	Depth	Dual	AP_{50}	AP_{75}	Params	FLOPs
✓	✗	✗	81.2	46.3	18.5 M	39.2 G
✗	✓	✗	74.5	41.8	18.5 M	39.2 G
✓	✓	✗	82.4	47.5	18.6 M	39.4 G
✓	✓	✓	**83.5**	**48.6**	18.8 M	39.8 G

The best performance metrics are highlighted in bold.

Depth modality, indicating that texture and color information in RGB images are crucial for underwater object classification and recognition. However, single-modality models show clear performance limitations in complex underwater scenarios.

The simple bimodal fusion approach, based on early channel concatenation of RGB and Depth, yields only a 1.2% improvement in AP50. This is primarily because early fusion introduces interference between modality-specific features, preventing the effective exploitation of the complementary characteristics of bimodal data. In contrast, the dual-backbone parallel feature extraction architecture proposed in this study achieves a 2.3% improvement in AP50. This design employs two independent branches to separately extract appearance features from the RGB modality and spatial features from the Depth modality, effectively avoiding early-stage cross-modal interference. It preserves the intrinsic advantages of both modalities and lays a solid foundation for subsequent fine-grained cross-modal fusion and feature enhancement.

4.4.2 Ablation study on the effectiveness of core modules

To validate the independent contributions and synergistic effects of the three core modules proposed in this study: UFEM (Underwater Feature Enhancement Module), UDFF (Underwater Denoising Feature Fusion Module), and D26Head (denoising detection head)—ablation experiments are conducted using the dual-backbone parallel architecture as the baseline under controlled settings. The experimental results are presented in Table 2.

From the results, it can be observed that incorporating the UFEM module alone improves the model's AP_{50} by 1.0% and AP_{75} by 1.1%, without introducing additional parameters or computational overhead. This demonstrates that UFEM effectively enhances feature representations of underwater small objects and low-contrast targets while suppressing background noise caused by water scattering, thereby improving the model's ability to capture fine-grained details. When the UDFF module is introduced independently, the model achieves a 1.6% improvement in AP_{50} and a 1.7% increase in AP_{75}, with only a marginal overhead of 0.1M parameters and 0.3G FLOPs. This indicates that UDFF enables adaptive complementary fusion between RGB texture information and depth structural information, effectively mitigating fusion interference caused by modality heterogeneity, and serving as a key component for accuracy improvement. Replacing the detection head with D26Head alone yields a 0.8% gain in AP_{50} and a 0.9% gain in AP_{75}, effectively alleviating the optimization conflict between classification and regression tasks under challenging underwater conditions such as blur and occlusion, thereby further improving localization precision and classification accuracy.

When all three modules are jointly applied, the model achieves 86.7% AP_{50} and 52.8% AP_{75}, representing a substantial 3.2% improvement in AP_{50} over the baseline. Notably, the combined performance gain exceeds the sum of individual improvements, demonstrating strong complementarity and synergistic effects among the modules. From the perspectives of feature enhancement, cross-modal fusion, and detection optimization, the three modules collectively enhance underwater object detection performance. The complete model contains 19.0 M parameters and 40.7 GFLOPs, representing only a slight increase of 0.2 M parameters and 0.9 GFLOPs compared to the baseline, achieving significant accuracy gains while maintaining a favorable balance between performance and model efficiency.

4.4.3 Internal ablation of the UDFF module

The UDFF module consists of four key components: Gaussian denoising, adaptive thresholding, crossmodal channel attention, and spatial attention. To quantify the contribution of each component, we conducted ablation experiments starting from the baseline channel concatenation fusion strategy. The experimental results are presented in Table 3. Gaussian denoising alone improves AP_{50} by 0.5% and AP_{75} by 0.5% without introducing any additional parameters or computational overhead. This demonstrates that underwater depth maps often contain substantial high-frequency noise caused by suspended particles, light scattering, and depth degradation. Simple Gaussian smoothing can effectively suppress such local noise, thereby purifying depth features and improving the quality of RGB-D feature fusion. Adaptive thresholding further improves AP_{50} by 0.4% and AP_{75} by 0.4%. This improvement indicates that underwater depth images may contain severe outlier regions and abnormal depth responses, which cannot be completely removed by Gaussian filtering alone. The adaptive thresholding strategy effectively suppresses these outliers while preserving important structural and edge information in the depth map. Cross-modal channel

TABLE 2 Ablation analysis on the effectiveness of core modules.

UFEM	UDFF	D26Head	AP_{50}	AP_{75}	Params	FLOPs
✗	✗	✗	83.5	48.6	18.8 M	39.8 G
✓	✗	✗	84.5	49.7	18.8 M	39.8 G
✗	✓	✗	85.1	50.3	18.9 M	40.1 G
✗	✗	✓	84.3	49.5	18.9 M	40.0 G
✓	✓	✓	**86.7**	**52.8**	19.0 M	40.7 G

The best performance metrics are highlighted in bold.

TABLE 3 Internal ablation analysis of the UDFF module.

GD	AT	CMA	SA	AP_{50}	AP_{75}	Params	FLOPs
✗	✗	✗	✗	83.5	48.6	18.8 M	39.8 G
✓	✗	✗	✗	84.0	49.1	18.8 M	39.8 G
✓	✓	✗	✗	84.4	49.5	18.8 M	39.8 G
✓	✓	✓	✗	84.9	50.1	18.9 M	40.0 G
✓	✓	✓	✓	**85.1**	**50.3**	18.9 M	40.1 G

The best performance metrics are highlighted in bold. GD denotes Gaussian Denoising. AT indicates Adaptive Thresholding. CMA presents Cross-Modal Channel Attention. SA denotes Spatial Attention.

attention brings 0.5% AP_{50} and 0.6% AP_{75} improvement with only a marginal increase in computational cost. This verifies that adaptive feature fusion based on channel correlation is far more effective than simple channel concatenation, as it can dynamically exploit the complementarity between RGB texture information and depth geometric cues. Spatial attention provides an additional 0.2% AP_{50} and 0.2% AP_{75} improvement. This component enhances the model's perception of local salient regions in underwater scenes, which is particularly beneficial for detecting small, blurred, and low-contrast targets.

To systematically evaluate the contribution of each component in the UDFF module and specifically validate the effectiveness of the clamp operation, we conducted internal ablation experiments. As shown in Table 4, the baseline was a dual-backbone architecture using simple channel-concatenation fusion, which achieved an AP_{50} of 83.5%.

The results show that each component of the UDFF module incrementally improved overall performance. Gaussian denoising alone improved AP_{50} by 0.5% by suppressing high-frequency random noise in the depth maps. Adding adaptive thresholding further improved AP_{50} by 0.7% by removing non-Gaussian outlier noise that could not be eliminated by Gaussian filtering alone. Incorporating the clamp operation provided additional improvements of 0.4% in AP_{50} and 0.5% in AP_{75}, confirming its important role in robust cross-modal fusion.

This performance gain supports our design rationale: in underwater environments, negative correlations between RGB and depth features are more likely to arise from noise, artifacts, and modality misalignment than from genuine complementary information. By truncating these negative correlations to zero, the clamp operation suppresses noise-induced conflicting signals while preserving positive complementary relationships between RGB texture and depth geometry. This design improves the reliability of the crossmodal attention mechanism, enabling the model to better exploit the complementarity between the two modalities in highly degraded underwater scenes.

4.4.4 Internal ablation of the UFEM module

The UFEM module is built upon two core designs: multi-scale depthwise separable convolutions and the Efficient Multi-scale Attention (EMA) mechanism. To evaluate their independent contributions, we conducted ablation experiments by replacing conventional convolutions in the neck network with different components. The experimental results are presented in Table 5. Multi-scale depthwise separable convolutions improve AP_{50} by 0.5% and AP_{75} by 0.6% without increasing parameters or computational complexity. This is because depthwise separable convolutions can capture richer multi-scale contextual information than standard convolutions, which is crucial for detecting underwater targets of different sizes. EMA mechanism alone achieves a 0.7% AP_{50} and 0.8% AP_{75} improvement. The EMA mechanism computes channel attention and multi-scale spatial attention in parallel, which effectively enhances the discriminative representation of low-contrast targets and suppresses background noise caused by water scattering. When both components are combined, the performance improvement exceeds the sum of individual improvements, demonstrating strong synergistic effects. The multi-scale convolutions provide rich local features, while the EMA mechanism adaptively weights these features to highlight salient target regions.

4.4.5 Ablation study on different attention mechanisms

To further evaluate the effectiveness of the proposed attention mechanism, we conducted a comparative ablation study against several widely adopted attention modules, including SE, CBAM, ECA, and CA, under the same dual-backbone parallel architecture as the baseline.

As shown in Table 6, the baseline model without an attention mechanism achieves an AP_{50} of 83.5% and an $AP_s$of 39.2%. All attention mechanisms consistently improve performance over the baseline, demonstrating the general benefit of attention-based feature refinement for underwater object detection. Specifically, SE and ECA provide moderate gains, increasing AP_{50} to 84.1% and 84.2%, respectively, while introducing negligible additional computational cost. CBAM and CA achieve slightly higher performance, with CA reaching an AP_{50} of 84.4% and an $AP_s$of 42.6%. This result indicates that spatial positional encoding is particularly beneficial for detecting small underwater targets. The proposed

TABLE 4 Ablation analysis on the clamp operation.

Gaussian denoising	Adaptive thresholding	Clamp operation	AP_{50}	AP_{75}	Params	FLOPs
✗	✗	✗	83.5	48.6	18.8 M	39.8 G
✓	✗	✗	84.0	49.1	18.8 M	39.8 G
✓	✓	✗	84.7	49.8	18.9 M	40.1 G
✓	✓	✓	**85.1**	**50.3**	18.9 M	40.1 G

The best performance metrics are highlighted in bold.

TABLE 5 Internal ablation analysis of the UFEM module.

Multi-scale DWConv	EMA	AP_{50}	AP_{75}	Params	FLOPs
✗	✗	83.5	48.6	18.8 M	39.8 G
✓	✗	84.0	49.2	18.8 M	39.8 G
✗	✓	84.2	49.4	18.8 M	39.8 G
✓	✓	**84.5**	**49.7**	18.8 M	39.8 G

The best performance metrics are highlighted in bold.

EMS attention module achieves the best overall performance, attaining an AP_{50} of 84.5% and an $AP_s$of 43.8%. These values represent improvements of 1.0 and 4.6 percentage points over the baseline, respectively.

Notably, EMS achieves these gains without increasing the number of parameters or computational complexity relative to the baseline, maintaining 18.8 M parameters and 39.8 GFLOPs. This result indicates a favorable balance between accuracy improvement and model efficiency. The superior performance in small-object detection suggests that EMS effectively enhances fine-grained feature representations, which are critical for detecting small and occluded targets in complex underwater scenes. These results demonstrate that the proposed EMS attention module is more effective than existing attention mechanisms as a feature refinement component for underwater object detection.

4.4.6 Internal ablation of the D26Head module

The D26Head module introduces two key improvements over the original YOLO26 head: Gaussian denoising preprocessing and the dual-branch training mechanism. To verify their effectiveness, we conducted ablation experiments by removing each component sequentially. The experimental results are presented in Table 7. Gaussian denoising preprocessing improves AP_{50} by 0.4% and AP_{75} by 0.4%. This component suppresses high-frequency noise in the feature maps introduced by underwater environments, providing cleaner inputs for the detection head and reducing false detections. One-to-many branch is critical for model training. Removing this branch leads to a significant performance drop of 0.6% AP_{50} and 0.6% AP_{75}. This is because the one-to-many branch provides abundant gradient signals during training, which helps the network quickly learn the general features of underwater targets and accelerates convergence. Oneto-one branch is essential for end-to-end inference. Removing this branch results in a severe performance degradation of 1.8% AP_{50} and 1.7% AP_{75}. This branch learns de-duplication capability through Hungarian matching during training, eliminating the need for NMS post-processing and significantly reducing missed detections in densely occluded scenes.

TABLE 7 Internal ablation analysis of the D26Head module.

GD	OTM	OTO	AP_{50}	AP_{75}	Params	FLOPs
✗	✓	✓	83.9	49.1	18.9 M	40.0 G
✓	✗	✓	83.7	48.9	18.9 M	40.0 G
✓	✓	✗	82.5	47.8	18.9 M	40.0 G
✓	✓	✓	**84.3**	**49.5**	18.9 M	40.0 G

The best performance metrics are highlighted in bold. GD denotes Gaussian Denoising. OTM indicates One-to-many branch. OTO presents One-to-one branch.

TABLE 6 Ablation results of different attention mechanisms.

Attention mechanism	AP_{50}	AP_s	Params (M)	GFLOPs
Baseline	83.5	39.2	18.8	39.8
SE Zhong et al. (2020)	84.1	41.5	18.8	39.9
CBAM Woo et al. (2018)	84.3	42.1	18.9	40.0
ECA Wang et al. (2020)	84.2	41.8	18.8	39.7
CA Hou et al. (2021)	84.4	42.6	18.9	40.1
EMS (Ours)	**84.5**	**43.8**	18.8	39.8

The best performance metrics are highlighted in bold.

4.4.7 Ablation study on cross-modal fusion strategies

To validate the effectiveness of the P_4/P_5 dual-level mid-stage fusion strategy adopted in this study, detection performance under different fusion positions and levels is compared. The experimental results are presented in Table 8.

The results demonstrate that the proposed P_4/P_5 dual-level mid-stage fusion strategy achieves the best detection performance. Early-stage fusion leads to severe degradation of modality-specific shallow features, resulting in an AP_{50} of only 82.3%, which is significantly lower than that of other fusion strategies. Fusion at a single level fails to fully exploit the semantic complementarity of bimodal data across different scales, providing limited improvement for underwater small object detection. Among these strategies, fusion at the P_4 level outperforms fusion at the P_5 level, indicating that mid-level features are more suitable for cross-modal fusion in underwater scenarios.

In contrast, the proposed P_4/P_5 dual-level mid-stage fusion strategy not only avoids the loss of shallow detail features caused by early fusion, but also leverages the complementary characteristics of bimodal data across multiple semantic scales through fusion at two deeper levels. With only a marginal increase in computational overhead, this strategy achieves the highest improvement in detection accuracy, thereby demonstrating its effectiveness and superiority.

4.4.8 Ablation study on robustness in extreme scenarios

To validate the effectiveness of the proposed core modules under adverse underwater conditions, the test set is divided into three typical sub-scenarios: clear water, heavy turbidity, and low-light deep water. Detection performance of the baseline model and the full model is evaluated separately across these scenarios. The experimental results are presented in Table 9.

From the results, it can be observed that the baseline dual-backbone model suffers from severe performance degradation under extreme underwater conditions: its AP_{50} drops from 85.3% in clear water to only 61.3% in heavy turbidity and 58.7% in low-

TABLE 8 Ablation analysis on cross-modal fusion strategies.

P_1/P_2	P_4	P_5	P_4/P_5	AP_{50}	Params	FLOPs
✓	✗	✗	✗	82.3	18.7 M	39.5 G
✗	✓	✗	✗	85.2	18.9 M	40.1 G
✗	✗	✓	✗	84.8	18.9 M	40.1 G
✗	✗	✗	✓	**86.7**	19.0 M	40.7 G

The best performance metrics are highlighted in bold.

TABLE 9 Ablation analysis on robustness in extreme scenarios.

Method	Clear water AP_{50}	Heavy turbidity AP_{50}	Low-light deep water AP_{50}
Baseline	85.3	61.3	58.7
None UFEM	86.1	72.4	70.3
None UDFF	86.5	74.1	72.6
UW-DualDet	**88.2**	**79.6**	**75.3**

The best performance metrics are highlighted in bold.

light deep water. After adding the proposed core modules, the complete UW-DualDet model achieves significant performance improvements across all scenarios. Specifically, it achieves an AP_{50} of 88.2% in clear water, 79.6% in heavy turbidity, and 75.3% in low-light deep water, representing improvements of 2.9%, 18.3%, and 16.6% respectively compared to the baseline model. Removing the UDFF module results in a 5.5% drop in AP_{50} in the heavy turbidity scenario, while removing the UFEM module leads to a 5.0% decrease in AP_{50} in the low-light deep water scenario, demonstrating the critical role of these modules in enhancing the model's robustness to extreme underwater imaging conditions.

4.4.9 Performance analysis of UW-DualDet

To validate the advancement, robustness and applicability of the proposed UW-DualDet, this paper conducts systematic fair comparisons with mainstream SOTA methods across general-purpose singlemodality, underwater-specific single-modality and underwater bimodal fusion models, with all experiments under identical configurations to ensure fairness.

As shown in Table 10, UW-DualDet achieves comprehensive leading performance on all detection accuracy metrics, with a core AP_{50} of 86.7%, and outperforms all competing methods on AP_{75} and $AP_{50:95}$. This demonstrates that the proposed bimodal architecture, UFEM and UDFF effectively improve underwater target feature discriminability and localization accuracy, and fully exploit the complementary characteristics of RGB and depth modalities. As shown by the heatmap rendering in Figure 9, in terms of target perception capability under dim and turbid backgrounds, UW-DualDet can localize underwater targets more accurately than YOLOv11 and UMIS-YOLO.

Compared with general-purpose single-modality models, UW-DualDet achieves 10.5%, 4.4%, 3.9% and 3.6% AP_{50} improvements over Faster R-CNN, YOLOv13, DEIM and Define, respectively. The core advantage of our bimodal design is that it compensates for the information loss of RGB modality in harsh underwater environments via depth structural information, breaking the performance ceiling of singlemodality methods that rely solely on easily degraded RGB appearance features. For underwater-specific single-modality models, UW-DualDet outperforms Boosting R-CNN, UW-YOLOv8 and FMSPP by 7.2%, 3.2% and 2.5% in AP_{50} respectively, with a larger leading margin on stricter metrics. It also surpasses the underwater bimodal models UMIS-YOLO(v12) and ICAFusion by 1.9% and 1.4% in AP_{50}, verifying that our lightweight modules have higher feature fusion efficiency than existing bimodal methods. The proposed UW-DualDet achieves an F1-score of 86.9%, which is 2.2% higher than the best-performing comparison method ICAFusion. This indicates that our method achieves a better balance between false positive and false negative detections, effectively reducing both missed detections and false alarms in complex underwater scenes.

TABLE 10 Comparative experimental analysis of performance against SOTA models.

Method	AP_{50}	AP_{75}	$AP50:95$	Precision	Recall	F1-scores
Faster R-CNN Girshick (2015)	76.2	42.5	39.6	78.5	72.3	75.3
YOLOv11s Khanam and Hussain (2024)	81.2	46.3	46.3	82.4	78.6	80.5
YOLOv12s Tian et al. (2025)	81.7	47.1	46.8	83.1	79.2	81.1
YOLOv13s Lei et al. (2025)	82.3	47.6	47.2	83.5	79.8.	81.6
DEIM-s Huang et al. (2025)	82.8	48.3	47.8	84.2	80.3	82.2
D-fine-s Peng et al. (2024)	83.1	48.5	48.1	84.1	80.6	82.5
Boosting R-CNN Song et al. (2023)	79.5	44.2	41.3	81.3	76.5	78.8
UW-YOLOv8 Guo et al. (2024)	83.5	48.7	48.5	84.8	81.2	83.0
FMSPP Hua et al. (2023b)	84.2	49.6	49.2	85.3	81.8	83.5
UMIS-YOLO Yang et al. (2025)	84.8	50.3	50.1	85.9	82.5	84.2
ICAFusion Shen et al. (2024)	85.3	51.2	50.7	86.4	83.1	84.7
UW-DualDet	**86.7**	**52.8**	**52.8**	**88.2**	**85.6**	**86.9**

The best performance metrics are highlighted in bold.

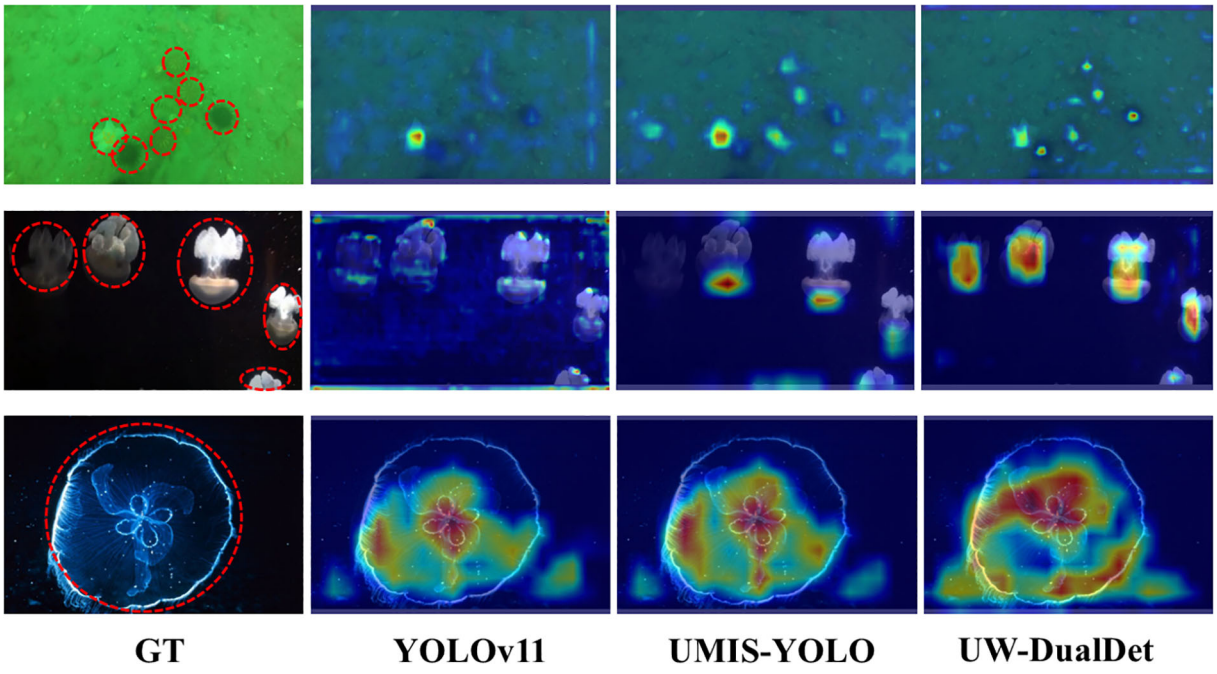

FIGURE 9
Comparative visualization of heatmap rendering perception.

4.4.10 Special robustness verification in complex underwater scenarios

To validate the detection stability of the proposed method in real-world complex underwater environments, this study conducts comparative experiments across three typical underwater scenarios: clear water, heavy turbidity, and low-light deep water. The AP_{50} results of different algorithms under each scenario are presented in Table 11.

In the clear-water scenario, all algorithms maintain relatively strong detection performance. The proposed UW-DualDet achieves an AP_{50} of 88.2%, slightly outperforming the best-performing comparison method, UMIS-YOLO, demonstrating robust baseline detection capability under favorable conditions. As illustrated in Figure 10 and Figure 11, UW-DualDet can accurately localize underwater targets at various scales in bright underwater environments.

In contrast, under extreme underwater conditions such as heavy turbidity and low-light deep water, singlemodality detection algorithms exhibit significant performance degradation. Conventional single-modality methods achieve an average AP_{50} of only 62.0% in heavy turbidity and 60.7% in low-light deep water. Even underwater-specific optimized single-modality models reach only 65.3% and 62.2% AP_{50}, respectively, under these challenging conditions. By comparison, the proposed UW-DualDet achieves an AP_{50} of 79.6% in heavy turbidity, outperforming the best competing method

TABLE 11 Comparative analysis of performance in complex underwater scenes.

Method	Clear water AP_{50}	Heavy turbidity AP_{50}	Low-light deep water AP_{50}
Faster R-CNN Girshick (2015)	80.5	58.3	55.2
YOLOv11s Khanam and Hussain (2024)	85.3	61.3	58.7
YOLOv12s Tian et al. (2025)	85.8	62.6	69.5
YOLOv13s Lei et al. (2025)	86.1	63.7	60.3
DEIM-s Huang et al. (2025)	86.0	64.1	60.5
D-fine-s Peng et al. (2024)	85.9	65.2	61.2
Boosting R-CNN Song et al. (2023)	83.2	60.5	57.4
UW-YOLOv8 Guo et al. (2024)	86.2	64.5	62.2
FMSPP Hua et al. (2023b)	86.3	65.3	60.3
UMIS-YOLO Yang et al. (2025)	87.9	72.4	71.2
ICAFusion Shen et al. (2024)	86.5	74.4	72.6
UW-DualDet	**88.2**	**79.6**	**75.3**

The best performance metrics are highlighted in bold.

FIGURE 10
Detection performance of UW-DualDet in bright underwater environments.

(ICAFusion) by 5.2 percentage points. In the low-light deep water scenario, it achieves an AP_{50} of 75.3%, exceeding ICAFusion by 2.7% AP_{50}. The proposed method consistently outperforms all competing algorithms across both extreme scenarios. According to the detection results presented in Figure 12 and Figure 13, UW-DualDet can still maintain robust detection performance under extreme conditions.

These results demonstrate that the proposed RGB–Depth bimodal architecture effectively compensates for the loss of texture information in the RGB modality under turbid and low-light

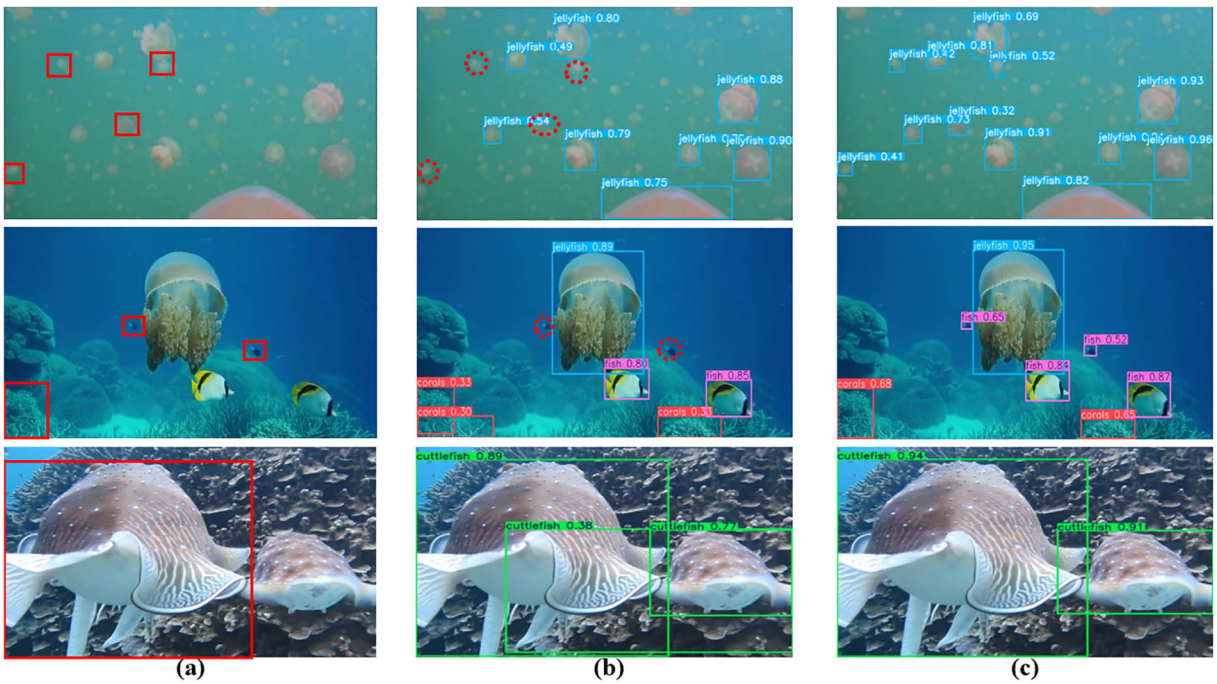

FIGURE 11
Performance comparison of the model in bright underwater environments. **(a)** indicates the missed and false detections of the baseline model, **(b)** shows the detection results of the baseline model, and **(c)** presents the detection results of UW-DualDet.

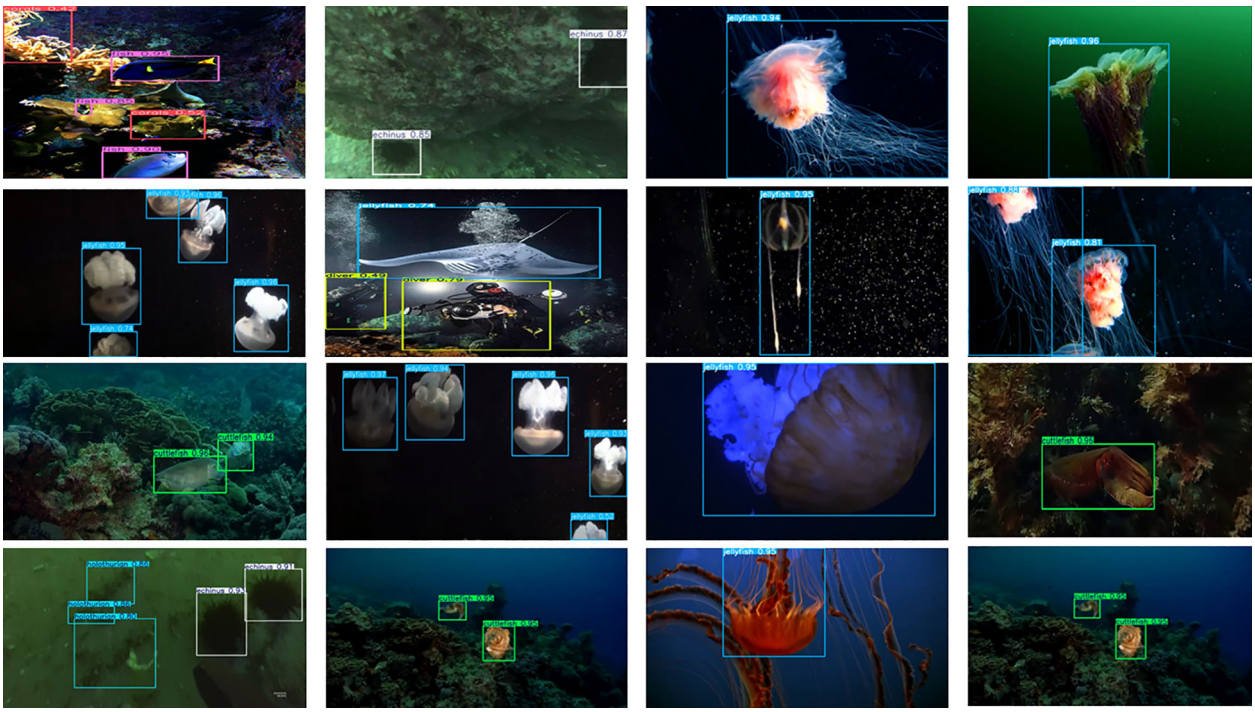

FIGURE 12
Detection performance of UW-DualDet in dim underwater environments.

conditions by leveraging structural information from the depth modality. Combined with feature enhancement and denoising fusion modules, the model significantly improves detection stability and environmental robustness in complex underwater scenarios.

4.4.11 Multi-scale underwater object detection performance analysis

To evaluate the adaptability of the proposed method to underwater targets of different scales, this study follows the standard

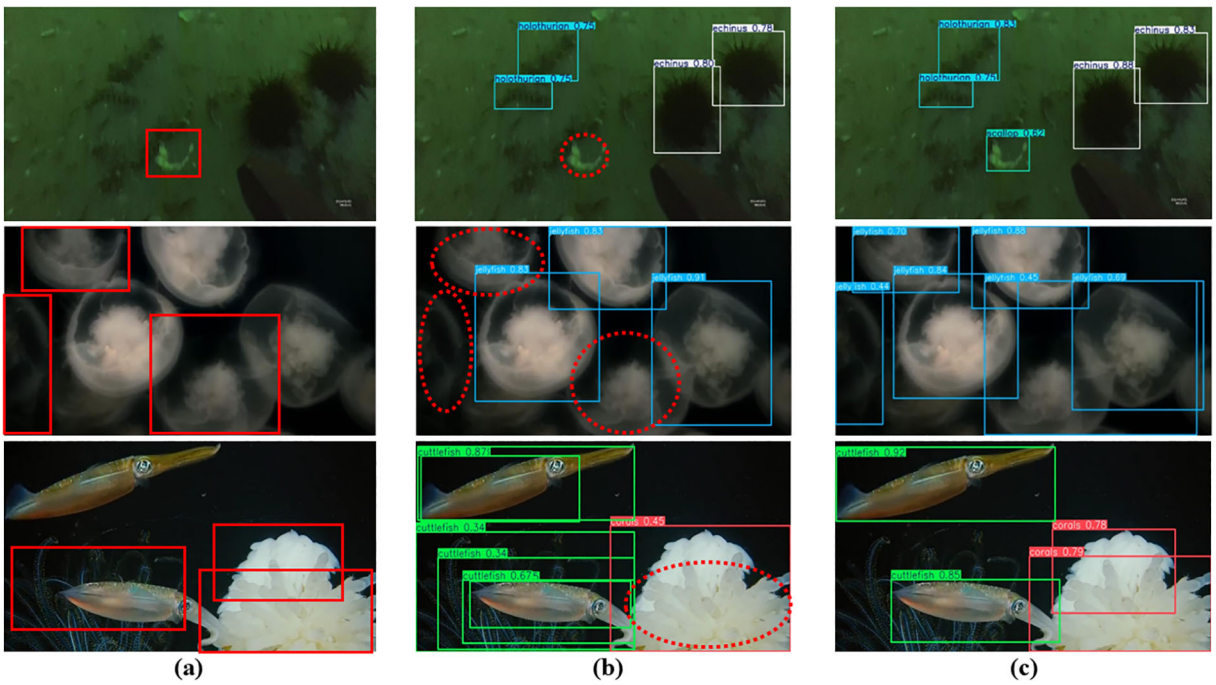

FIGURE 13
Performance comparison of the model in dim underwater environments. **(a)** indicates the missed and false detections of the baseline model, **(b)** shows the detection results of the baseline model, and **(c)** presents the detection results of UW-DualDet.

TABLE 12 Comparative analysis of multi-scale underwater object detection performance.

Method	AP_S	AP_M	AP_L
YOLOv11s Khanam and Hussain (2024)	35.2	62.4	79.5
YOLOv12s Tian et al. (2025)	35.8	62.7	79.8
YOLOv13s Lei et al. (2025)	35.5	62.5	79.6
DEIM-s Huang et al. (2025)	38.2	64.5	81.2
Dfine-s Peng et al. (2024)	39.1	65.2	81.5
Boosting R-CNN Song et al. (2023)	34.5	60.3	78.6
UW-YOLOv8 Guo et al. (2024)	41.3	66.8	82.4
FMSPP Hua et al. (2023b)	43.5	68.2	83.1
UMIS-YOLO Yang et al. (2025)	45.2	70.4	84.6
ICAFusion Shen et al. (2024)	46.7	71.6	85.2
UW-DualDet	**47.9**	**72.3**	**86.3**

The best performance metrics are highlighted in bold.

definitions of the COCO dataset and divides test objects into three categories: small objects (pixel area$< 32^2$, AP_S), medium objects ($32^2 \leq$ pixel area$< 96^2$, AP_M), and large objects (pixel area $\geq 96^2$, AP_L). Comparative experiments are conducted with mainstream algorithms, and the results are presented in Table 12.

From the experimental results, it can be observed that lightweight models in the YOLO series do not exhibit a linear performance improvement trend with newer versions in underwater scenarios. YOLOv12s shows only marginal improvement over YOLOv11s, while YOLOv13s even demonstrates a slight performance decline. This indicates that general detection optimizations designed for natural scenes provide limited gains in underwater environments characterized by turbidity and low contrast, and cannot effectively adapt to the unique properties of underwater targets.

In terms of scale-specific performance, all algorithms maintain stable performance for large object detection, with relatively small differences among models. The proposed UW-DualDet achieves an AP_L of 86.3%, which is only 1.1% higher than the second-best model, ICAFusion, indicating that most methods can reliably detect large-scale underwater targets. For medium-sized objects, the proposed method achieves an AP_M of 72.3%, outperforming the best competing model, ICAFusion, by 0.7%, and exceeding general single-modality models by nearly 10% on average, demonstrating a consistent performance advantage. For small object detection, which represents a core challenge in underwater scenarios, the proposed method exhibits the most significant advantage. It achieves an AP_S of 47.9%, outperforming the best-performing general single-modality model (Define-s) by 8.8%, the best underwater-specific single-modality model (FMSPP) by 4.4%, and the best bimodal model (ICAFusion) by 1.2%. Furthermore, as presented in Figures 14, 15, shows that UW-DualDet can maintain stable detection performance for densely occluded targets as well as underwater targets at different scales.

These results demonstrate that the proposed RGB–Depth bimodal architecture and feature enhancement design effectively leverage spatial contour information from the depth modality to compensate for weak texture representation of small underwater targets in RGB images, which are often obscured by background noise. This significantly enhances small-object feature representation while maintaining strong detection performance across multiple scales, thereby addressing the core challenge of multi-scale underwater object detection.

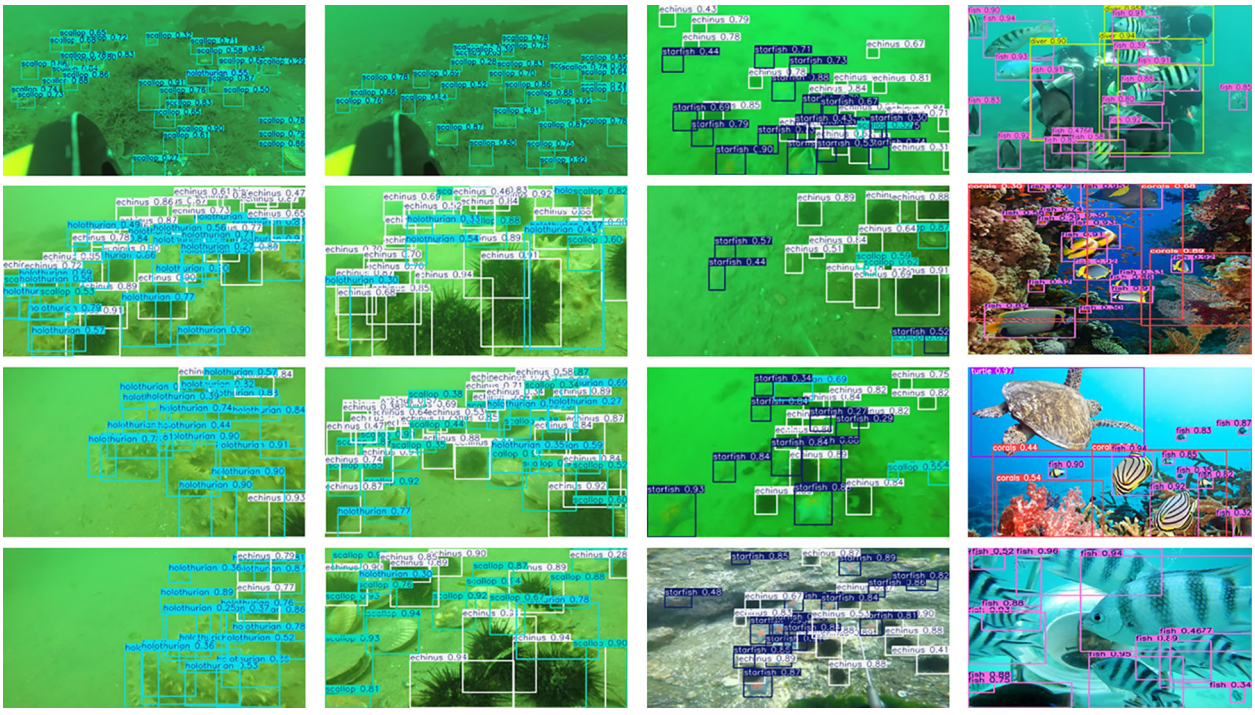

FIGURE 14
Detection performance of UW-DualDet for densely occluded underwater targets.

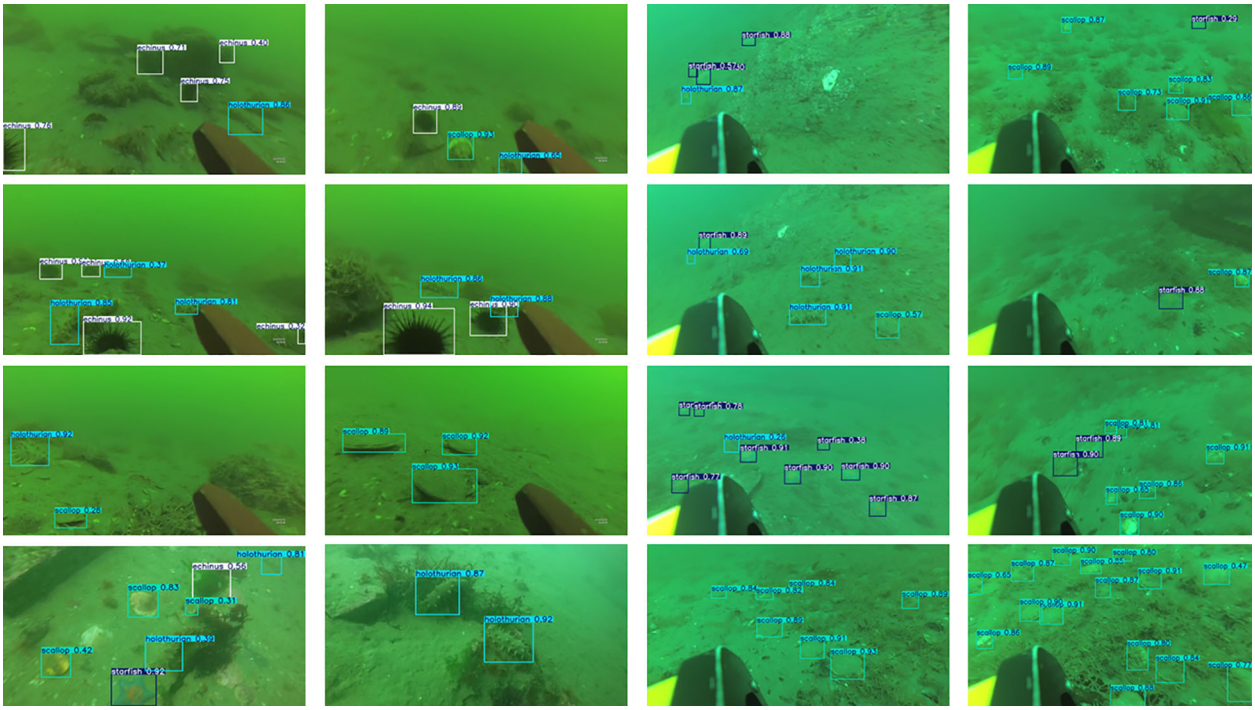

FIGURE 15
Detection performance of UW-DualDet for multi-scale targets in turbid environments.

4.4.12 Model complexity and real-time performance analysis

To validate the deployment feasibility of the proposed UW-DualDet and objectively analyze real-time performance differences between underwater multimodal and single-modality detection models, this study adopts floating-point operations (GFLOPs) and parameter count (Params) as metrics for model complexity. Frames per second (FPS), measured under a unified input resolution of 640 × 640 with single-batch inference, is used as the metric for real-time performance. All evaluations are conducted under a consistent hardware environment. The comparison results with mainstream single-modality and multimodal detection methods are presented in Table 13.

From the comparison results, a common pattern in underwater multimodal detection can be observed: under the same baseline architecture, multimodal detection models generally exhibit significantly lower inference speed than single-modality models. Taking different versions of the YOLO series as examples, the single-modality YOLOv8m achieves an inference speed of 128 FPS, whereas its multimodal counterpart, UMIS-YOLOv8m, increases GFLOPs from 110 to 130 and parameters from 27M to 39M due to dual-input processing and cross-modal fusion, resulting in a reduced FPS of 83, corresponding to a 35.2% decrease in real-time performance. Similarly, YOLOv11m achieves 132 FPS in the single-modality setting, whereas UMIS-YOLOv11m drops to 72 FPS, a 45.5% decrease. For YOLOv12m, the single-modality model reaches 108 FPS, while UMIS-YOLOv12m achieves only 56 FPS, corresponding to a 48.1% reduction. This phenomenon arises from the inherent requirement of multimodal models to process both RGB and depth inputs simultaneously, where additional dual-branch feature extraction and cross-modal fusion modules inevitably introduce extra computational overhead, significantly reducing inference speed and constituting a core bottleneck in current underwater multimodal detection research. As an RGB–Depth multimodal detection model, the proposed UW-DualDet follows this general trend. Its inference speed of 116 FPS shows a 12.1% gap compared to the single-modality YOLOv11m (22M parameters, 132 FPS) of similar scale, primarily due to the inherent computational cost of dual-modality inputs. However, compared to existing mainstream multimodal detection models, UW-DualDet achieves a substantial improvement in lightweight design and real-time performance. In terms of model complexity, UW-DualDet contains only 19M parameters, which is approximately 46%–49% of the UMIS series multimodal models, and requires only 40.7 GFLOPs, corresponding to 26.6%–31.3% of their computational cost, while also being significantly lower than all compared single-modality models. In terms of real-time performance, the 116 FPS of UW-DualDet is 1.40×, 1.61×, and 2.07× that of UMIS-YOLOv8m, UMIS-YOLOv11m, and UMIS-YOLOv12m, respectively, and even surpasses the inference speed of single-modality YOLOv12m and ASF-YOLO.

As shown in Table 14, all single-modal detection methods only report the inference performance of the detection model itself, excluding the depth map generation step. In contrast, the performance metrics of all multi-modal methods cover the complete end-to-end pipeline from RGB image input to final detection result output. The experimental results demonstrate that UW-DualDet achieves comprehensive leading real-time performance among multi-modal detection methods. Its end-to-end inference latency

TABLE 13 Comparative analysis of model real-Time performance and computational complexity.

Method	GFLOPs	FPS	Params
YOLOv8m Kukartsev et al. (2024)	110	128	27M
YOLOv11m Khanam and Hussain (2024)	123	132	22M
YOLOv12m Tian et al. (2025)	123	108	22M
ASF-YOLO Kang et al. (2024)	155	47	48M
UMIS-YOLOv8m Yang et al. (2025)	130	83	39M
UMIS-YOLOv11m Yang et al. (2025)	150	72	40M
UMIS-YOLOv12m Yang et al. (2025)	153	56	41M
UW-DualDet	**40.7**	**116**	**19M**

The best performance metrics are highlighted in bold.

is only 8.6 ms/frame, which is 28.3% lower than that of the best-performing comparison method UMIS-YOLOv8m. The end-to-end FPS reaches 98, representing a 36.1% improvement over UMIS-YOLOv8m. Furthermore, the peak GPU memory consumption of UW-DualDet is only 2.4 GB, which is 36.8%–42.9% lower than that of the UMIS series models. Most importantly, the inference latency of UW-DualDet is very close to that of single-modal YOLOv8m (7.8 ms) and YOLOv11m (7.6 ms), effectively alleviating the common trade-off in multi-modal detection where accuracy improvements are accompanied by a dramatic increase in computational complexity.

As shown in Table 15, the results show that the proposed modules introduce minimal additional computational overhead, with 0.2M additional parameters and 0.9G additional GFLOPs, and 1.0ms additional per-frame latency. This fully verifies the lightweight, plug-and-play characteristics of the proposed modules, which can be easily integrated into other detection frameworks with negligible performance loss.

TABLE 14 Comparative analysis of model real-time performance.

Method	Inference latency	End-to-end FPS	Peak GPU memory (GB)
YOLOv8m Kukartsev et al. (2024)	7.8	–	3.2
YOLOv11m Khanam and Hussain (2024)	7.6	–	3.0
YOLOv12m Tian et al. (2025)	9.3	–	3.1
ASF-YOLO Kang et al. (2024)	21.3	–	4.5
UMIS-YOLOv8m Yang et al. (2025)	12.0	72	3.8
UMIS-YOLOv11m Yang et al. (2025)	13.9	63	4.0
UMIS-YOLOv12m Yang et al. (2025)	17.9	49	4.2
UW-DualDet	8.6	98	2.4

End-to-End FPS: Frames per second measured for the complete end-to-end pipeline, including both monocular depth map generation using the DepthAnything-V2-Large model and subsequent object detection inference. Peak GPU Memory (GB): Maximum GPU memory consumption during the entire inference process. Inference Latency (ms/frame): Total end-to-end inference latency per image, measured as the time elapsed from inputting an RGB image to obtaining the final detection results.

TABLE 15 Module-level computational overhead.

Module	Additional params (M)	Additional GFLOPs	Additional latency
UDFF	0.1	0.2	0.3
UFEM	0.0	0.0	0.2
D26Head	0.1	0.7	0.5
Total	0.2	0.9	1.0

Combined with the previously reported accuracy results, UW-DualDet not only significantly outperforms all competing methods in detection accuracy, but also effectively alleviates the common trade-off in underwater multimodal detection, where accuracy improvements are typically accompanied by substantial increases in computational complexity and reductions in real-time performance. The proposed method narrows the real-time performance gap between multimodal and single-modality models, achieving a strong balance between detection performance and deployment efficiency in real-world underwater scenarios. The limitations of this work and a comprehensive discussion of future research directions are presented in Supplementary Section S2 of the Supplementary Materials.

5 Conclusion

This paper proposes an underwater bimodal denoising detection network to address key challenges in complex underwater environments, such as image degradation and feature suppression due to illumination fluctuations and water scattering, as well as high missed-detection rates and limited scene robustness caused by dense target occlusion. The proposed framework is specifically designed to address three critical issues in underwater object detection: feature enhancement, multimodal denoising fusion, and occluded target detection.

Specifically, the method consists of three core modules. First, the Underwater Feature Enhancement Module (UFEM) is introduced to mitigate feature degradation caused by illumination fluctuations and water scattering. UFEM incorporates multi-scale feature interaction, illumination-adaptive correction, and background noise suppression to effectively enhance the discriminative representation of targets under challenging lighting conditions. Second, the Underwater Denoising Feature Fusion (UDFF) module is designed to address residual intra-modal noise, severe cross-modal interference, and underutilization of complementary information. It performs adaptive alignment and fusion of RGB texture-color cues with depth-based spatial-structural information, fully leveraging bimodal complementarity. Third, the occlusion-aware detection head, D26Head, is introduced to improve detection performance in densely occluded scenes. By incorporating spatial contour priors from the depth modality and constructing a residual feature mining branch for occluded targets, it significantly enhances detection accuracy and recall in dense occlusion scenarios.

Extensive ablation and comparative experiments on a self-constructed underwater RGB-Depth object detection dataset

demonstrate the effectiveness of the proposed network. The method achieves an mAP_{50} of 86.7%, outperforming mainstream single-modal and multimodal state-of-the-art approaches. In representative illumination-challenging scenarios, including low-light deep water and highly turbid conditions, the network attains AP_{50} scores of 75.3% and 79.6%, respectively. It also achieves an AP_{50} of 88.2% in clear water conditions, which includes overexposure sub-scenarios where the model maintains consistent performance, demonstrating strong robustness and environmental adaptability across diverse underwater imaging conditions. In dense occlusion scenarios, the network achieves an AP_{50} of 72.4% for occluded targets, surpassing the best comparable method by 3.6%. In addition, the network operates at 116 FPS, meeting the real-time requirements of practical applications such as underwater intelligent inspection and robotic operations.

Data availability statement

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

Author contributions

CZ: Conceptualization, Data curation, Investigation, Methodology, Writing – original draft, Writing – review & editing. XL: Conceptualization, Methodology, Writing – review & editing. SW: Conceptualization, Investigation, Methodology, Software, Supervision, Writing – review & editing. ZZ: Investigation, Software, Visualization, Writing – review & editing. XC: Data curation, Formal analysis, Funding acquisition, Resources, Validation, Writing – review & editing. TX: Data curation, Investigation, Project administration, Validation, Writing – review & editing. WY: Investigation, Methodology, Software, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This research was supported by Youth Program of the Shandong Provincial Natural Science Foundation: Spatiotemporal Processes and Mechanisms of Land-Use Urbanization Transformation Based on GIS and Remote Sensing(ZR2022QD146).

Acknowledgments

We express our gratitude to the Weihai Institute of Marine Information Science and Technology for providing the computational resources for model training.

Conflict of interest

The author(s) declared that this work 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) declared that generative AI was not 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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmars.2026.1859313/full#supplementary-material

References

Cai, W., Zhu, J., and Zhang, M. (2025). Multi-modality object detection with sonar and underwater camera via object-shadow feature generation and saliency information. *Expert Syst. Appl.* 287, 128021. doi: 10.1016/j.eswa.2025.128021

Cao, Y., Bin, J., Hamari, J., Blasch, E., and Liu, Z. (2023). "Multimodal object detection by channel switching and spatial attention", in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada: IEEE 403–411.

Chen, G., Mao, Z., Tu, Q., and Shen, J. (2024a). A cooperative training framework for underwater object detection on a clearer view. *IEEE Trans. Geosci. Remote Sens.* 62, 1–17. doi: 10.1109/tgrs.2024.3440386

Chen, L., Zhou, F., Wang, S., Dong, J., Li, N., Ma, H., et al. (2022). Swipenet: Object detection in noisy underwater scenes. *Pattern Recognit.* 132, 108926. doi: 10.1016/j.patcog.2022.108926

Chen, Y., Wang, B., Zhu, W., and Yuan, J. (2024b). "Rgb-ir yolo combining modality-specific reconstruction and information integration", in: *2024 39th Youth Academic Annual Conference of Chinese Association of Automation (YAC)* (Dalian, China: IEEE), 2045–2050. doi: 10.1109/YAC63405.2024.10598725

Chen, X., Zeng, Q., Wei, C., Cheng, G., and Li, X. (2026). A unified underwater object detection network with image adaptive enhancement and attention-driven feature fusion. *Optics Laser Technol.* 199, 115028. doi: 10.1016/j.optlastec.2026.115028

Fu, C., Liu, R., Fan, X., Chen, P., Fu, H., Yuan, W., et al. (2023). Rethinking general underwater object detection: Datasets, challenges, and solutions. *Neurocomputing* 517, 243–256. doi: 10.1016/j.neucom.2022.10.039

Gao, J., Zhang, Y., Geng, X., Tang, H., and Bhatti, U. A. (2024). Pe-transformer: Path enhanced transformer for improving underwater object detection. *Expert Syst. Appl.* 246, 123253. doi: 10.1016/j.eswa.2024.123253

Girshick, R. (2015). "Fast r-cnn", in: *Proceedings of the IEEE international conference on computer vision*, Boston, Massachusetts, USA: IEEE 1440–1448.

Guo, A., Sun, K., and Zhang, Z. (2024). A lightweight yolov8 integrating fasternet for real-time underwater object detection. *J. Real-Time Image Process.* 21, 49. doi: 10.1007/s11554-024-01431-x

Han, L., Liu, X., Xu, T., and Wang, Y. (2025). Hybrid mamba for amphibious limulidae low-light image enhancement. *Front. Mar. Sci.* 12, 1578735. doi: 10.3389/fmars.2025.1578735

Hou, Q., Zhou, D., and Feng, J. (2021). "Coordinate attention for efficient mobile network design", in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Seattle, Washington, USA: IEEE 13713–13722.

Hua, X., Cui, X., Xu, X., Qiu, S., Liang, Y., Bao, X., et al. (2023a). Underwater object detection algorithm based on feature enhancement and progressive dynamic aggregation strategy. *Pattern Recognit.* 139, 109511. doi: 10.1016/j.patcog.2023.109511

Hua, X., Cui, X., Xu, X., Qiu, S., Liang, Y., Bao, X., et al. (2023b). Underwater object detection algorithm based on feature enhancement and progressive dynamic aggregation strategy. *Pattern Recognit.* 139, 109511. doi: 10.1016/j.patcog.2023.109511

Huang, S., Lu, Z., Cun, X., Yu, Y., Zhou, X., and Shen, X. (2025). "Deim: Detr with improved matching for fast convergence", in: *Proceedings of the computer vision and pattern recognition conference*, Nashville, Tennessee, USA: IEEE 15162–15171.

Jian, M., Yang, N., Tao, C., Zhi, H., and Luo, H. (2024). Underwater object detection and datasets: a survey. *Intelligent Mar. Technol. Syst.* 2, 9. doi: 10.1007/s44295-024-00023-6

Jin, H.-S., Cho, H., Jiafeng, H., Lee, J.-H., Kim, M.-J., Jeong, S.-K., et al. (2022). Hovering control of uuv through underwater object detection based on deep learning. *Ocean Eng.* 253, 111321. doi: 10.1016/j.oceaneng.2022.111321

Jin, S., Li, G., Xu, Z., Zhang, Y., Qin, Z., Liu, H., et al. (2026). Hybrid attention triple branch transformer net for underwater image enhancement. *Pattern Recognit. Lett.* 201 (2026), 95–102. doi: 10.1016/j.patrec.2026.01.014

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

Khanam, R., and Hussain, M. (2024). Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*. doi: 10.48550/arXiv.2410.17725

Kotthapalli, M., Ravipati, D., and Bhatia, R. (2025). Yolov1 to yolov11: A comprehensive survey of real-time object detection innovations and challenges. *arXiv preprint arXiv:2508.02067*. doi: 10.48550/arXiv.2508.02067

Kukartsev, V., Ageev, R., Borodulin, A., Gantimurov, A., and Kleshko, I. (2024). "Deep learning for object detection in images development and evaluation of the yolov8 model using ultralytics and roboflow libraries", in: *Computer Science On-line Conference* (Cham: Springer), 629–637. doi: 10.1007/978-3-031-70285-3_48

Lei, M., Li, S., Wu, Y., Hu, H., Zhou, Y., Zheng, X., et al. (2025). Yolov13: Real-time object detection with hypergraph-enhanced adaptive visual perception. *arXiv preprint arXiv:2506.17733*. doi: 10.48550/arXiv.2506.17733

Li, B., Ma, Q., Zhu, Z., Deng, S., Hu, H., and Li, X. (2025a). Uaed-net: a unified adaptive enhancement and detection network with multi-scale feature refinement for underwater scenarios. *Meas. Sci. Technol.* 36, 105407. doi: 10.1088/1361-6501/ae09c5

Li, M., Pang, H., and Jiang, C. (2025b). Sgl-yolo: Lightweight underwater object detection algorithm based on feature fusion: M. li et al. *Signal. Image Video Process.* 19, 1372. doi: 10.21203/rs.3.rs-6510637/v1

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

Li, R., Xiang, J., Sun, F., Yuan, Y., Yuan, L., and Gou, S. (2023). Multiscale cross-modal homogeneity enhancement and confidence-aware fusion for multispectral pedestrian detection. *IEEE Trans. Multimedia* 26, 852–863. doi: 10.1109/tmm.2023.3272471

Liang, X., Zhang, T., Yu, H., and Liu, Z. (2025). Rcf-yolo: an underwater object detection algorithm based on improved yolov10n. *J. Real-Time Image Process.* 22, 97. doi: 10.1007/s11554-025-01677-z

Lv, C., and Pan, W. (2025). Luod-yolo: A lightweight underwater object detection model based on dynamic feature fusion, dual path rearrangement and cross-scale integration. *J. Real-Time Image Process.* 22, 204. doi: 10.1007/s11554-025-01778-9

Peng, Y., Li, H., Wu, P., Zhang, Y., Sun, X., and Wu, F. (2024). D-fine: Redefine regression task in detrs as fine-grained distribution refinement. *arXiv preprint arXiv:2410.13842*. 2020, 44015–44031. doi: 10.48550/arXiv.2410.13842

Rao, W., Chen, G., Zhang, Y., Cang, J., Chen, S., and Wang, C. (2025). Benthos-detr: a high-precision efficient network for benthic organisms detection. *Front. Mar. Sci.* 12, 1586510. doi: 10.3389/fmars.2025.1586510

Sapkota, R., Cheppally, R. H., Sharda, A., and Karkee, M. (2025). Yolo26: key architectural enhancements and performance benchmarking for real-time object detection. *arXiv preprint arXiv:2509.25164*. doi: 10.48550/arXiv.2509.25164

Shen, J., Chen, Y., Liu, Y., Zuo, X., Fan, H., and Yang, W. (2024). Icafusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognit.* 145, 109913. doi: 10.1016/j.patcog.2023.109913

Song, P., Li, P., Dai, L., Wang, T., and Chen, Z. (2023). Boosting r-cnn: Reweighting r-cnn samples by rpn's error for underwater object detection. *Neurocomputing* 530, 150–164. doi: 10.1016/j.neucom.2023.01.088

Sun, W., Liu, X., Hao, J., Yao, Q., Xi, H., Wu, Y., et al. (2025). Ags-yolo: An efficient underwater smallobject detection network for low-resource environments. *J. Mar. Sci. Eng.* 13, 1465. doi: 10.3390/jmse13081465

Tian, Y., Ye, Q., and Doermann, D. (2025). Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*. 78433–78457. doi: 10.48550/arXiv.2502.14740

Tian, Y., Zhang, Y., and Zhang, H. (2023). Recent advances in stochastic gradient descent in deep learning. *Mathematics* 11, 682. doi: 10.3390/math11030682

Tu, W., Yu, H., Wu, Z., Li, J., Cui, Z., Yang, Z., et al. (2025). Yolov10-udfishnet: detection of diseased takifugu rubripes juveniles in turbid underwater environments. *Aquacult. Int.* 33, 125. doi: 10.1007/s10499-024-01798-5

Wang, H., Sun, S., Bai, X., Wang, J., and Ren, P. (2023a). A reinforcement learning paradigm of configuring visual enhancement for object detection in underwater scenes. *IEEE J. Oceanic Eng.* 48, 443–461. doi: 10.1109/joe.2022.3226202

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., and Hu, Q. (2020). "Eca-net: Efficient channel attention for deep convolutional neural networks", in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Seattle, Washington, USA: IEEE 11534–11542.

Wang, Z., Yu, Z., and Zheng, B. (2025). Yolo-nerfslam: underwater object detection for the visual nerf-slam. *Front. Mar. Sci.* 12, 1582126. doi: 10.3389/fmars.2025.1582126

Wang, H., Zhong, G., Sun, J., Chen, Y., Zhao, Y., Li, S., et al. (2023b). Simultaneous restoration and super-resolution gan for underwater image enhancement. *Front. Mar. Sci.* 10, 1162295. doi: 10.3389/fmars.2023.1162295

Wei, W., Wei, M., Li, H., and Hong, G. (2026). A scene-adapted lightweight yolo-ld model for dead fish detection in recirculating aquaculture systems. *Aquacult. Int.* 34, 59. doi: 10.2139/ssrn.5705210

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). "Cbam: Convolutional block attention module", in: *Proceedings of the European conference on computer vision (ECCV)*, Munich, Germany: Springer 3–19.

Xu, H., He, J., Pang, C., and Yu, Y. (2025). Sfudnet: Underwater object detection via spatial-frequency domain modulation with mixture of experts. *Knowledge-Based Syst.* 324, 113805. doi: 10.1016/j.knosys.2025.113805

Yang, Y., Feng, X., Li, M., Hu, X., Qin, J., Gruen, A., et al. (2025). Umis-yolo: Underwater multimodal images instance segmentation with yolo. *IEEE Trans. Geosci. Remote Sens.* doi: 10.1109/tgrs.2025.3618269

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., et al. (2024). Depth anything v2. *Adv. Neural Inf. Process. Syst.* 37, 21875–21911. doi: 10.52202/079017-0688

Yang, S., Zhang, X., and Tan, P. (2026). Fs2-detr: Transformer-based few-shot sonar object detection with enhanced feature perception. *J. Mar. Sci. Eng.* 14, 304. doi: 10.3390/jmse14030304

Ye, Z., Peng, X., Li, D., and Shi, F. (2025). Yolov11-mse: A multi-scale dilated attention-enhanced lightweight network for efficient real-time underwater target detection. *J. Mar. Sci. Eng.* 13, 1843. doi: 10.3390/jmse13101843

Yin, H., Cheng, Y., and Shi, W. (2026). Tfdf-yolo: A position detection model for underwater wireless power transfer docking. *J. Mar. Sci. Eng.* 14, 429. doi: 10.3390/jmse14050429

Yu, M., Xie, Z., Ye, Y., and Shi, C. (2025). Tmae-yolo: precision detection of mud crabs in underwater environments. *Aquacult. Int.* 33, 462. doi: 10.1007/s10499-025-02149-8

Zhang, H., Fan, S., Zou, S., Yu, Z., and Zheng, B. (2024). Deep underwater image compression for enhanced machine vision applications. *Front. Mar. Sci.* 11, 1411527. doi: 10.3389/fmars.2024.1411527

Zhang, J., and Gao, B. (2025). Rcdi-yolo: a target-detection method for complex environment side-scan sonar images based on improved yolov8. *Front. Mar. Sci.* 12, 1679077. doi: 10.3389/fmars.2025.1679077

Zhao, Y., Fan, Z., Li, H., Zhang, R., Xiang, W., Wang, S., et al. (2023). Symmetricnet: end-to-end mesoscale eddy detection with multi-modal data fusion. *Front. Mar. Sci.* 10, 1174818. doi: 10.3389/fmars.2023.1174818

Zhao, Z., Liu, Y., Sun, X., Liu, J., Yang, X., and Zhou, C. (2021). Composited fishnet: fish detection and species recognition from low-quality underwater videos. *IEEE Trans. Image Process.* 30, 4719–4734. doi: 10.1109/tip.2021.3074738

Zhao, G., Zhang, K., Wang, L., Zhao, W., and Zhang, W. (2025). Cidnet: cross-scale interference mining detection network for underwater object detection. *Knowledge-Based Syst.* 324, 113902. doi: 10.1016/j.knosys.2025.113902

Zhong, Z., Lin, Z. Q., Bidart, R., Hu, X., Daya, I. B., Li, Z., et al. (2020). "Squeeze-and-attention networks for semantic segmentation", in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, Washington, USA: IEEE 13065–13074.

Zhou, W., Zeng, J., Li, S., and Zhang, Y. (2025). Uw-yolo-bio: a real-time lightweight detector for underwater biological perception with global and regional context awareness. *J. Mar. Sci. Eng.* 13, 2189. doi: 10.3390/jmse13112189