

Article

UW-YOLO-Bio: A Real-Time Lightweight Detector for Underwater Biological Perception with Global and Regional Context Awareness

Wenhao Zhou ^{1,2} , Junbao Zeng ^{1,*}, Shuo Li ¹ and Yuexing Zhang ¹ ¹ Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China; zhouwenhao@sia.cn (W.Z.); shuoli@sia.cn (S.L.); zhangyuexing@sia.cn (Y.Z.)² University of Chinese Academy of Sciences, Beijing 100049, China

* Correspondence: zengjb@sia.cn

Abstract

Accurate biological detection is crucial for autonomous navigation of underwater robots, yet severely challenged by optical degradation and scale variation in marine environments. While image enhancement and domain adaptation methods offer some mitigation, they often operate as disjointed preprocessing steps, potentially introducing artifacts and compromising downstream detection performance. Furthermore, existing architectures struggle to balance accuracy, computational efficiency, and robustness across the extreme scale variability of marine organisms in challenging underwater conditions. To overcome these limitations, we propose UW-YOLO-Bio, a novel framework built upon the YOLOv8 architecture. Our approach integrates three dedicated modules: (1) The Global Context 3D Perception Module (GCPM), which captures long-range dependencies to mitigate occlusion and noise without the quadratic cost of self-attention; (2) The Channel-Aggregation Efficient Downsampling Block (CAEDB), which preserves critical information from low-contrast targets during spatial reduction; (3) The Regional Context Feature Pyramid Network (RCFPN), which optimizes multi-scale fusion with contextual awareness for small marine organisms. Extensive evaluations on DUO, RUOD, and URPC datasets demonstrate state-of-the-art performance, achieving an average improvement in mAP50 of up to 2.0% across benchmarks while simultaneously reducing model parameters by 8.3%. Notably, it maintains a real-time inference speed of 61.8 FPS, rendering it highly suitable for deployment on resource-constrained autonomous underwater vehicles (AUVs).

Keywords: underwater object detection; multi-scale feature fusion; real-time inference

Academic Editor: Fabio Bruno

Received: 22 October 2025

Revised: 15 November 2025

Accepted: 17 November 2025

Published: 18 November 2025

Citation: Zhou, W.; Zeng, J.; Li, S.; Zhang, Y. UW-YOLO-Bio: A Real-Time Lightweight Detector for Underwater Biological Perception with Global and Regional Context Awareness. *J. Mar. Sci. Eng.* **2025**, *13*, 2189. <https://doi.org/10.3390/jmse13112189>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Embodied intelligence is reshaping the autonomous obstacle avoidance capabilities of underwater robots in complex marine environments, with accurate biological detection serving as a fundamental prerequisite. In deep-sea exploration, the ability of an underwater robot to merely “detect” obstacles without being able to “recognize” their categories imposes severe limitations on its autonomous navigation. For example, the failure to differentiate between a “movable fish” and a “stationary obstacle requiring evasion” can lead to non-specific avoidance strategies, resulting in unnecessary path deviations or increased risk of collision.

As the core of autonomous navigation systems, underwater object detection directly determines the reliability of obstacle avoidance decisions. In contrast to terrestrial appli-

cations, where state-of-the-art detectors such as YOLOv8 [1], DETR [2], and RT-DETR [3] have achieved performance approaching that of human perception, underwater vision is significantly compromised by multiple degradation factors: wavelength-dependent light attenuation introduces severe color distortion and loss of chromatic information; forward and backward scattering by suspended particles produces haze-like blurring and reduced contrast; dynamic illumination and non-uniform lighting obscure object boundaries. Collectively, these factors reduce feature discriminability, disrupt spatial coherence, and substantially increase false detection rates, ultimately undermining the overall performance of underwater obstacle avoidance systems.

Early efforts to mitigate underwater optical degradation predominantly focused on image restoration and enhancement. Physics-based approaches, such as the Jaffe–McGlamery model [4], attempt to model the light transport process; however, their reliance on precise optical parameters, which are difficult to measure in practice, constrains their applicability. Deep learning-based methods (e.g., WaterNet [5], UWCNN [6]) trained on large-scale synthetic datasets have demonstrated improved color correction, but they often suffer from domain shift, resulting in poor generalization to real-world conditions. More recently, self-supervised and unsupervised techniques have gained attention. For instance, FUnIE-GAN [7] leverages generative adversarial training without paired data, while Water-Color Transfer [8] exploits color mapping strategies. Although these methods show promising enhancement capabilities, they are typically deployed as standalone preprocessing steps, disconnected from the downstream detection task. This separation prevents end-to-end optimization and, in some cases, introduces artifacts or excessive smoothing that negatively impact final detection accuracy.

To address issues of low contrast and noise, research has also emphasized the development of robust feature extractors. CNN-based models such as LMFEN [9], which employs hierarchical convolutions to retain fine-grained features, and UW-ResNet [10], which enhances blurred boundaries through modified residual structures, have achieved significant results. The integration of attention mechanisms has further improved feature discriminability: CBAM [11,12] and SE blocks [13] recalibrate features spatially and channel-wise; lightweight modules such as those used in DAMO-YOLO [14] enhance detection accuracy without compromising efficiency. Despite these advances, the suitability of Transformer-based architectures for underwater tasks remains contested. Although the Swin Transformer [15] demonstrates strong capabilities in modeling long-range dependencies, its quadratic computational complexity and high memory cost limit its deployment on resource-constrained embedded platforms. Lightweight alternatives such as MobileViT [16] and EfficientFormer [17] offer efficiency advantages but often underperform in detecting small underwater targets.

The scale variability of marine organisms further complicates underwater object detection. Multi-scale architectures based on feature pyramids have therefore become a standard solution. The Feature Pyramid Network (FPN) [18] improves multi-scale performance by integrating high-level semantics with low-level details via a top-down pathway. Extensions such as PANet [19] incorporate a bottom-up pathway to strengthen cross-scale feature propagation, while BiFPN [20] adaptively adjusts scale contributions using weighted bidirectional fusion. Nevertheless, these general strategies encounter unique challenges underwater. Shallow features are often contaminated by scattering and suspended particles, resulting in high noise and blurred details. Directly fusing such degraded features with semantic-rich high-level representations introduces interference and degrades detection performance. Recent approaches, such as UW-FPN [21], selectively enhance features at different pyramid levels or denoise shallow maps prior to fusion.

However, these solutions typically incur higher computational costs and demonstrate limited adaptability across diverse underwater environments.

To overcome these limitations, we propose Underwater YOLO-based Biological detection method (UW-YOLO-Bio), a high-precision framework for underwater biological detection that significantly improves the autonomy of underwater robots. The framework aims to enhance both detection accuracy and robustness, thereby ensuring reliable input for subsequent obstacle avoidance decisions. Our method achieves notable improvements on the DUO, RUOD, and URPC datasets, increasing mAP50 by 2% and reduce model parameters by 8%.

Specifically, UW-YOLO-Bio introduces three innovations tailored to underwater detection challenges:

- (1) Global Context Perception Module (GCPM): Enhances global feature extraction, effectively mitigating underwater optical degradation and improving perception robustness.
- (2) Channel-Aggregation Efficient Downsampling Block (CAEDB): Preserves critical information during spatial reduction, thereby improving sensitivity to low-contrast underwater targets.
- (3) Regional Context Feature Pyramid Network (RCFPN): Optimizes multi-scale feature fusion through contextual awareness, substantially improving detection accuracy for small marine organisms.

Beyond performance gains, this work contributes a systematic framework for environment-aware detector design. By translating domain-specific knowledge—such as underwater optical physics and marine ecological characteristics—into architectural innovations, UW-YOLO-Bio provides a generalizable design paradigm. We anticipate that the principles underlying this framework can inspire the development of robust visual systems in other adverse environments, including foggy driving conditions, smoke-obscured disaster scenarios, and deep-space imaging. Furthermore, the principles underlying UW-YOLO-Bio, particularly its robustness to noise and clutter and its capability for multi-scale target detection, demonstrate potential applicability beyond optical imagery. For instance, in the field of maritime surveillance, High-Frequency Surface Wave Radar (HFSWR) is pivotal for over-the-horizon target detection. Similar to underwater optical environments, HFSWR data, especially Range-Doppler (RD) maps, are contaminated by strong sea clutter and noise, making target detection challenging [22–24]. The GCPM ability to model global context could help distinguish weak target signatures from clutter, while the RCFPN multi-scale fusion might adapt to targets of varying velocities and sizes on the RD map. We believe an adapted version of our method could contribute to improved target detection and clutter suppression in HFSWR systems, and we regard this as a promising direction for future interdisciplinary research.

The remainder of this paper is organized as follows. Section 2 reviews related work on underwater object detection, multi-scale feature fusion, and downsampling methods. Section 3 details the proposed UW-YOLO-Bio framework, including GCPM, CAEDB, and RCFPN. Section 4 presents the experimental setup, datasets, ablation studies, comparative results, and an analysis of failure cases. Finally, Section 5 concludes the paper and discusses future work.

2. Related Works

2.1. Underwater Object Detection

As a fundamental technology for marine exploration, underwater object detection has garnered increasing research attention in recent years.

2.1.1. Image Enhancement Preprocessing Methods

Early studies primarily focused on image enhancement as a preprocessing stage, aiming to improve detection accuracy by enhancing the visual quality of raw underwater imagery. Ji et al. [25] introduced a collaborative framework integrating enhancement and detection, reporting an mAP50 increase of 2.1% (from 78.4% to 80.5%). However, its reliance on precise optical models restricts generalizability across diverse aquatic environments. Wang et al. [26] proposed a Retinex-based method that decouples illumination and reflection components to recover details, but its high computational cost hinders real-time deployment. Similarly, Zhang et al. [27] utilized physics-based modeling of absorption and scattering to enhance image contrast, yet the method demonstrated limited adaptability to complex real-world underwater scenarios. Overall, these enhancement-driven approaches face inherent limitations in robustness, computational efficiency, and environmental adaptability, highlighting the need for more integrated detection frameworks.

2.1.2. Domain Adaptation

Domain adaptation has emerged as another promising direction, aiming to reduce the discrepancy between terrestrial and underwater domains. Dai et al. [28] proposed the Gated Cross-Domain Collaborative Network (GDCDN), which dynamically modulates cross-domain features through a gating mechanism, achieving an mAP50 of 81.7%. Nevertheless, the method incurs an 18% increase in GFLOPs, posing challenges for embedded deployment. Alternatively, Liu et al. [29] adopted adversarial learning to minimize inter-domain distribution gaps, achieving a 2.3% gain in mAP50. Despite these improvements, adversarial approaches are vulnerable to substantial performance degradation under highly variable underwater conditions. These drawbacks emphasize the necessity of designing domain adaptation strategies that are both computationally efficient and robust to diverse aquatic environments.

2.1.3. Dedicated Architectural Improvements

To address underwater-specific challenges, researchers have proposed architectural modifications tailored to this domain. For instance, LMFEN [9] employs Hierarchical Deep Convolution (HDC) to preserve fine-grained features during downsampling, reaching 81.3% mAP50. However, its limited modeling of global context reduces effectiveness in cluttered scenes with small objects. Recent advancements in object detection have seen the application of the latest YOLO architectures combined with transformer backbones for challenging real-world tasks. For instance, Seunghyeon et al. [30,31] demonstrated the efficacy of a Swin Transformer-based YOLOv10 model for multi-class non-PPE detection on construction sites, achieving a high average precision while maintaining real-time performance. CST-YOLO [32] combines CNNs with Swin Transformer to capture global dependencies via window-based self-attention, achieving 92.7% mAP50 on a blood cell dataset. Its significant computational overhead, however, (25% increase in GFLOPs) limits real-time feasibility. For resource-constrained platforms, lightweight designs remain a priority. Mobile-YOLO [33], which integrates a MobileNet backbone and lightweight detection head, reduces parameters by 32.2% and improves throughput by 95.2% (63.5 FPS), though it provides only a marginal accuracy gain of 0.8%. Similarly, CAEDB-YOLO [34] applies a Channel-Aggregation Efficient Downsampling Block, reaching 82.7% mAP50 with a 5.7% parameter reduction, but with limited improvements for small-object detection. Collectively, these approaches reveal a persistent trade-off between detection accuracy, computational efficiency, and robustness across varying object scales in underwater environments.

2.2. Multi-Scale Feature Fusion

Multi-scale feature fusion is a cornerstone of modern object detection, enabling recognition of objects across diverse size distributions. However, in underwater scenarios, the large size variability of marine organisms, combined with optical degradation-induced feature blurring, exacerbates the challenge. Classical multi-scale designs such as the Feature Pyramid Network (FPN) [18] exhibit reduced effectiveness in underwater tasks, achieving only 78.3% mAP50 on the DUO dataset [35], a 4.2% drop compared to terrestrial performance. Similarly, Spatial Pyramid Pooling (SPP) [36] and its variants, such as SPPCSPC [37], enhance feature aggregation and achieve 82.7% mAP50. Nonetheless, pooling operations inherently discard fine spatial details critical for small-object detection. On the DUO dataset, SPPCSPC small-object performance drops to 75.8% [35], 6.3% lower than in terrestrial environments. Alternative approaches, such as channel-split fusion (e.g., MCS [33]), which processes features across channels with different receptive fields, have demonstrated competitive results in other domains (91.3% mAP50 on blood cell data). Yet, their application to underwater detection remains limited, with only a marginal 0.9% gain observed on the DUO dataset. This gap highlights the urgent need for more effective and adaptive multi-scale feature fusion techniques in underwater contexts.

2.3. Downsampling Methods

Downsampling plays a critical role in determining both the accuracy and efficiency of object detectors. Conventional operations, including max pooling and strided convolution, are widely adopted but face severe limitations underwater. For instance, MPConv [38], which relies on max pooling, suffers substantial feature loss in low-contrast underwater imagery, resulting in an mAP50 of 76.2% on the DUO dataset—an 8.5% decrease relative to terrestrial conditions. Depthwise separable convolution, as used in MobileNet [39], reduces parameters by 75%, but offers only marginal improvements (0.5% mAP50 increase) in underwater detection and fails to preserve critical details of small objects. Channel-aggregation approaches have been introduced to mitigate these drawbacks. CAEDB [40], for example, enhances representational power by aggregating multi-channel information, achieving 82.7% mAP50 with a 5.7% reduction in parameters—representing a 6.5% improvement over traditional downsampling on the DUO dataset. Nonetheless, CAEDB does not explicitly account for underwater optical properties. To address this, Zhang et al. [41] proposed an adaptive method that dynamically adjusts downsampling parameters based on absorption and scattering characteristics, attaining 83.2% mAP50. However, its dependence on accurate optical measurements, which are often difficult to obtain, restricts its practicality across diverse environments. This limitation points to the need for more robust and adaptive downsampling strategies that can generalize across heterogeneous underwater conditions.

3. Methodology

This section presents the architecture of the proposed Underwater YOLO-based Biological detection (UW-YOLO-Bio) framework, which is designed around three core principles: environmental adaptability, semantic awareness, and computational efficiency. YOLOv8 was selected as the baseline architecture for this study due to its proven stability, extensive documentation, and optimal balance between accuracy and speed for resource-constrained platforms at the time of our experimental design. Its modular structure also facilitated the seamless integration of our proposed modules. The overall architecture of the proposed UW-YOLO-Bio is depicted in Figure 1; the framework enhances the YOLOv8 backbone by integrating three novel modules:

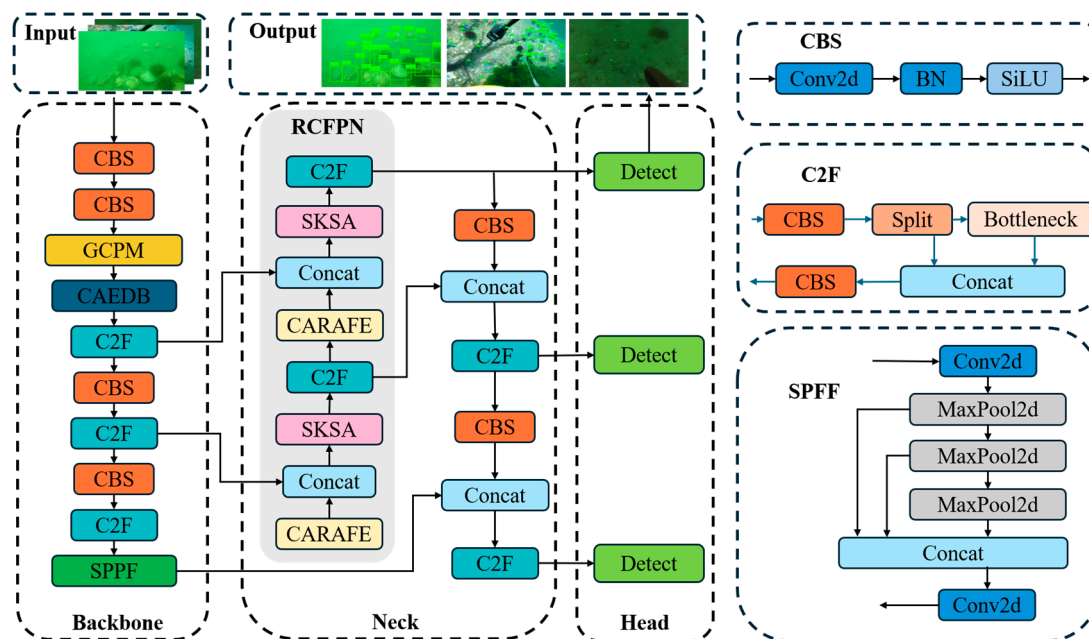


Figure 1. Overall framework diagram of UW-YOLO Bio.

- (1) GCPM—efficiently captures long-range dependencies without the quadratic computational cost of conventional self-attention, thereby providing robust resistance to underwater occlusion and noise;
- (2) CAEDB—utilizes depthwise separable convolutions and channel aggregation to achieve efficient feature reduction while retaining critical information from low-contrast targets;
- (3) RCFPN—refines multi-scale feature fusion by embedding regional contextual cues and ecological scale priors, leading to substantial improvements in detecting small marine organisms.

The GCPM, CAEDB, and RCFPN modules are not isolated components but form a synergistic pipeline addressing distinct bottlenecks in underwater detection. The GCPM acts as a feature enhancer within the backbone network. By capturing global contextual dependencies, it mitigates the macroscopic effects of occlusion and noise, thereby providing a cleaner and semantically richer feature foundation for subsequent modules. The CAEDB operates at critical downsampling stages. Its efficient channel aggregation mechanism ensures that the information crucial for low-contrast targets, which is enhanced by the GCPM, is preserved during spatial reduction, preventing the loss of fine details during transmission. Finally, the RCFPN receives and fuses these optimized multi-scale features from the preceding modules. Its regional context awareness and content-aware upsampling capability enable precise reconstruction of target details, particularly for small objects that have been effectively highlighted by the GCPM and CAEDB. In essence, the GCPM is responsible for macro-environment understanding and robustness, the CAEDB guarantees the lossless transmission of critical information, and the RCFPN focuses on the accurate localization and recognition of multi-scale targets. The three modules work in concert to collectively enhance the model robustness in complex underwater environments.

The overall architecture of the proposed UW-YOLO-Bio is depicted in Figure 1. The framework enhances the YOLOv8 backbone by integrating three novel modules. The input image is first processed by initial convolutional layers. The GCPMs are strategically inserted into the backbone to capture long-range dependencies and mitigate occlusion and noise at different scales. The feature maps then pass through a series of CAEDB modules,

which perform efficient spatial reduction while preserving critical information from low-contrast targets. The multi-scale features extracted by the backbone are then fed into the RCFPN, which optimizes feature fusion through its Separable Kernel Spatial Attention (SKSA) and Content-Aware ReAssembly of Features (CARAFE) components, enhancing the representation for objects of various sizes, particularly small marine organisms. Finally, the detection head (on the right) utilizes these refined, multi-scale features to predict bounding boxes and class probabilities. These components operate synergistically to address the unique challenges inherent in underwater biological detection.

3.1. Global Context 3D Perception Module

GCPM serves as a pivotal component of UW-YOLO-Bio, specifically engineered to capture long-range dependencies in underwater object detection. Unlike traditional approaches relying on computationally intensive self-attention mechanisms, GCPM adopts a lightweight yet powerful design integrating three core elements: a C2f module, a Global Context Block (GCBlock), and a channel-wise SimAM mechanism, as shown in Figure 2. This integration enables efficient global context modeling while maintaining minimal computational overhead.

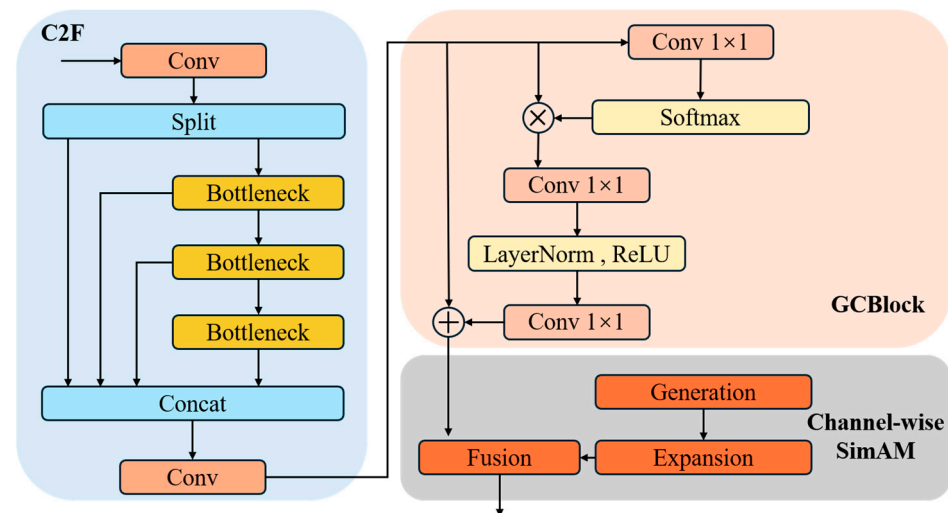


Figure 2. Global Context 3D Perception Module (GCPM) Structure diagram.

The GCPM takes a feature map of size $C \times H \times W$ as input and produces an output feature map of the identical dimension $C \times H \times W$. While the spatial and channel dimensions remain unchanged, the representational capacity of the output features is significantly enhanced through the integration of global context and channel-wise attention.

3.1.1. Global Context Block (GCBlock)

At the heart of GCPM lies the GCBlock, which replaces conventional non-local blocks with a lightweight convolution-based alternative. As depicted in Figure 2, GCBlock follows a streamlined three-stage process:

Global Feature Extraction: Global average pooling followed by a 1×1 convolution compresses the input feature map into a compact context vector.

Feature Transformation: A two-layer bottleneck structure (1×1 convolution $\rightarrow$ LayerNorm $\rightarrow$ ReLU $\rightarrow 1 \times 1$ convolution) converts this context vector into a channel-specific weight map.

Feature Aggregation: The refined global context is reintroduced into the original features through element-wise addition.

Mathematically, the GCBlock operation can be expressed as Equation (1):

$$z_i = x_i + W_{v2} \text{ReLU} \left(\text{LN} \left(W_{v1} \sum_{j=1}^{N_p} \frac{e^{W_k x_j}}{\sum_{m=1}^{N_p} e^{W_k x_m}} x_j \right) \right) \quad (1)$$

where $\frac{e^{W_k x_j}}{\sum_{m=1}^{N_p} e^{W_k x_m}}$ is the weight for the global attention pool, and $W_{v2} \text{ReLU}(\text{LN}(W_{v1}(\cdot)))$ denotes the bottleneck transform.

The ability of GCBlock to capture long-range dependencies stems from its use of Global Average Pooling (GAP). The GAP operation aggregates information across the entire spatial dimension ($H \times W$) into a compact context vector. This operation effectively allows any single point in the feature map to establish a connection with all other points, regardless of their spatial distance. The subsequent bottleneck transformation (W_{v1} and W_{v2}) learns to weigh this global information and reintegrates it into the original features via an element-wise addition. This mechanism shares a similar spirit with the self-attention in Transformers, which models pairwise interactions. However, by performing spatial aggregation first, the GCBlock reduces the computational complexity from quadratic, $O((HW)^2)$, typical of standard self-attention, to linear, $O(HWC)$, thereby achieving efficient long-range dependency modeling suitable for resource-constrained platforms.

This design is critically important for underwater applications, as it achieves a substantial reduction in computational complexity, which adds a mere 0.2% more parameters, while fully preserving the module's capacity to capture global contextual information. Unlike standard self-attention mechanisms, whose computational cost scales quadratically with feature map size, the proposed GCBlock exhibits linear complexity with respect to input dimensionality. This efficiency makes the module exceptionally well-suited for deployment on computationally restricted underwater platforms.

3.1.2. Channel-Wise SimAM

Complementing the GCBlock, the Channel-wise SimAM module introduces 3D perception capabilities without additional learnable parameters. This parameter-free mechanism adaptively assigns importance weights to individual channels according to their intrinsic feature characteristics, thereby enhancing discrimination between target objects and background noise in underwater imagery.

The workflow of the SimAM, as illustrated in Figure 2, can be conceptually divided into three distinct stages—Generation, Expansion, and Fusion—which collectively enable efficient global context modeling:

Generation: This initial stage generates a compact global context vector from the input feature map X using Global Average Pooling (GAP). The GAP operation aggregates spatial information from each channel ($H \times W$) into a single scalar value, producing a channel-wise descriptor vector of size $C \times 1 \times 1$. This vector serves as a foundational summary of the global contextual information present in the feature map.

Expansion: The generated context vector then undergoes transformation and expansion within a bottleneck structure comprising two 1×1 convolutional layers separated by a LayerNorm and ReLU activation. This stage is designed to capture complex, non-linear interactions between channels. The first convolution potentially elevates the dimensionality to learn richer representations, while the second convolution projects the features back to the original channel dimension C . This process effectively recalibrates the channel-wise context, learning to emphasize informative features and suppress less useful ones.

Fusion: Finally, the transformed context vector is fused back into the original input features X via a broadcasted element-wise addition. This integration mechanism allows the refined global context to modulate features across all spatial locations, enhancing

the discriminative power of the network without altering the spatial dimensions of the feature map ($C \times H \times W$). This residual-style fusion ensures stable training and seamless integration of global information.

The design of Channel-wise SimAM is inspired by the spatial suppression mechanism in mammalian visual cortices, where active neurons inhibit nearby neuronal activity. This biological phenomenon is mathematically formulated as an energy function that measures the linear separability between a neuron and its neighboring neurons within the same channel. The energy function is defined as Equation (2):

$$e_t^* = \frac{4(\sigma^2 + \lambda)}{(t - \mu)^2 + 2\sigma^2 + 2\lambda} \quad (2)$$

where t represents the value of the target neuron, μ and σ^2 are the mean and variance of all neurons in the channel, and λ is a regularization parameter.

The importance of each neuron is inversely proportional to its energy value $1/e_t^*$, with lower energy indicating higher importance. The channel-wise SimAM module then computes the importance map as Equation (3):

$$E_{inv} = \frac{d}{4(v + \lambda)} + 0.5 \quad (3)$$

where $d = (X - \mu_X)^2$ and $v = \frac{\sum_{i=1}^H \sum_{j=1}^W d_{ij}}{n}$ with $n = H \times W - 1$ (spatial size of the feature map). Finally, the feature map is refined using Equation (4):

$$F_{out} = F_{in} \odot \text{sigmoid}(E_{inv}) \quad (4)$$

where $\odot$ denotes element-wise multiplication. The implementation of Channel-wise SimAM is remarkably simple, which demonstrates its parameter-free nature and computational efficiency. This simplicity is a key advantage over other attention modules that require extensive parameter tuning.

The GCPM integrates GCBlock and Channel-wise SimAM into a single efficient framework, with the following computational flow Equation (5):

$$F_{gcpm} = \text{Channel-wise SimAM}(\text{GCBlock}(F_{in})) \quad (5)$$

This integration strategy ensures that the global context information captured by GCBlock is further refined by Channel-wise SimAM, enhancing the model's ability to perceive contextual relationships in underwater environments.

3.1.3. Discussion on GCPM

The novelty of the proposed GCPM lies not in the invention of its sub-components but in their novel integration and the specific problem they are designed to address. The GCBlock is adapted from the work of Cao et al. [42], which efficiently captures long-range dependencies. The Channel-wise SimAM mechanism is inspired by the parameter-free attention module proposed by Yang et al. [43], which leverages neuroscience-based energy functions. Our key innovation is the synergistic combination of these two distinct mechanisms into a single, cohesive module (GCPM) tailored for underwater visual degradation. Specifically, the GCBlock first models broad contextual relationships to clean the features, and its output is subsequently refined by the SimAM mechanism, which enhances discriminative local features by suppressing neuronal noise. This specific wiring and the focus on mitigating underwater optical artifacts (e.g., haze, low contrast) through 3D perception represent the novel contribution of the GCPM, offering a more effective and efficient alternative to standard self-attention for our target domain.

To quantitatively substantiate the efficiency claims of the proposed GCBLOCK and provide a clear comparison with existing attention mechanisms, we conducted a detailed computational complexity analysis. This analysis aims to validate the theoretical advantages of linear complexity over quadratic approaches, such as standard self-attention, and to demonstrate the practical benefits in terms of floating-point operations (FLOPs) and parameter counts. Specifically, we evaluated the following key modules under a representative feature map size (e.g., a common intermediate resolution in underwater detection tasks):

Standard Self-Attention: Representing the baseline with quadratic complexity $O((H \times W)^2)$, which is computationally expensive for high-resolution feature maps.

GCBLOCK (Ours): Our proposed module, designed with linear complexity $O(H \times W \times C)$ through global average pooling and bottleneck transformations.

Channel-wise SimAM (Ours): The parameter-free component integrated into GCPM, also exhibiting linear complexity.

The results, summarized in Table 1, highlight the dramatic reduction in computational overhead achieved by our modules. For instance, the GCBLOCK reduces FLOPs by over $25\times$ compared to self-attention, while maintaining competitive performance. This analysis not only corroborates the theoretical foundations outlined in Section 3.1 but also underscores the practicality of our approach for resource-constrained deployments. By presenting these metrics, we provide empirical evidence that the GCBLOCK and SimAM collectively offer an optimal balance between efficiency and effectiveness, aligning with the goals of real-time underwater biological detection.

Table 1. Computational complexity analysis of attention mechanisms (for a representative feature map size).

Module	Theoretical Complexity	Flops (G)	Parameters (K)
Standard Self-Attention	$O((H \times W)^2)$	2.15	1024
GCBLOCK	$O(H \times W \times C)$	0.08	0.5
Channel-wise SimAM	$O(H \times W \times C)$	0.05	0

3.2. Channel Aggregation Efficient Downsampling Block

Downsampling serves as a pivotal operation in object detection networks, as it reduces spatial dimensions to alleviate computational burden while enlarging receptive fields. However, in underwater imaging scenarios characterized by low contrast, feature fragmentation, and high noise sensitivity, conventional downsampling techniques, such as max pooling and stride convolution, would often lead to considerable loss of discriminative information.

To overcome these limitations, we propose the CAEDB, designed to preserve crucial features while maintaining computational efficiency. As illustrated in Figure 3, the CAEDB consists of two principal components: Depthwise Separable Convolution (DWConv) and the Channel Aggregation Block (CABlock). This dual-structure design enhances representational capacity and stabilizes feature extraction in degraded underwater conditions through a complementary two-stage process. The DWConv first performs efficient spatial filtering, reducing noise and extracting preliminary features while expanding the receptive field. The subsequent CABlock then operates on these refined features, adaptively recalibrating channel-wise responses based on global context. This sequential processing allows the module to first mitigate spatial degradation (e.g., blurring, noise) and then address feature-level imbalances (e.g., low contrast between targets and background), leading to more robust and stable feature representations compared to using either component alone.

In contemporary architectures, channel mixing, which is denoted as CMixer($\cdot$), is commonly implemented via two linear projections. Typical realizations include a two-layer channel-wise multilayer perceptron (MLP) interleaved with a 3×3 DWConv. The kernel

size of 3×3 is chosen for the DWConv as it represents an optimal trade-off between receptive field size and computational efficiency. Smaller kernels (e.g., 1×1) lack sufficient spatial context for effective feature extraction in noisy underwater environments, while larger kernels (e.g., 5×5 or 7×7) significantly increase computational complexity and parameters without providing proportional performance gains for this task [39,40]. The depthwise separable design factorizes a standard convolution into a depthwise convolution (applying a single filter per input channel) followed by a pointwise convolution (1×1 convolution), which drastically reduces computational cost and model parameters while maintaining representative capacity, making it ideal for real-time applications on resource-constrained platforms. However, due to significant information redundancy among channels, standard MLP-based mixers often require a substantial expansion in the number of parameters (e.g., by a factor of 4 to 8 in the hidden layer) to achieve satisfactory performance. This parameter inflation directly leads to poor computational efficiency, characterized by increased FLOPs, larger memory footprint, and slower inference speeds, which are critical drawbacks for deployment on embedded systems.

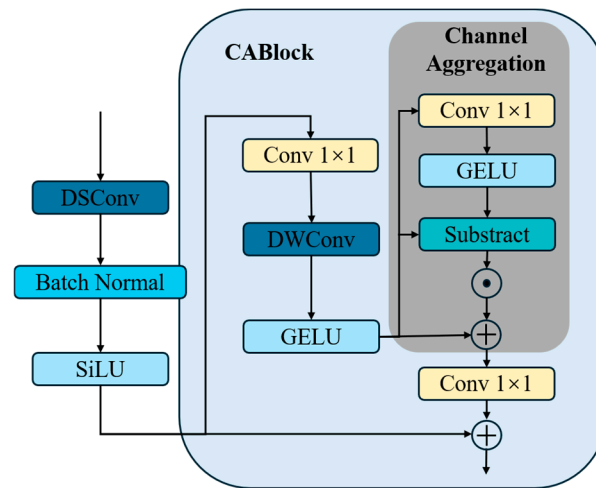


Figure 3. Channel Aggregation Efficient Downsampling Block (CAEDB) Structure diagram.

To alleviate this inefficiency, several recent studies have incorporated channel enhancement modules, such as the Squeeze-and-Excitation (SE) block, into the MLP pathway to recalibrate channel responses. Inspired by frequency-domain insights, recent research has introduced a lightweight Channel Aggregation (CA) module [40], which adaptively redistributes channel-wise feature representations within high-dimensional latent spaces—without introducing additional projection layers. The specific implementation process of this mechanism is depicted as Equations (6) and (7):

$$Y = \text{GELU} \left(\text{DW}_{3 \times 3} \left(\text{Conv}_{1 \times 1} \left(\text{Norm}(X) \right) \right) \right) \quad (6)$$

$$Z = \text{Conv}_{1 \times 1}(\text{CA}(Y)) + X \quad (7)$$

Concretely, DW represents depthwise convolution, $\text{CA}(\cdot)$ is implemented by a channel-reducing projection $W_r : \mathbb{R}^{C \times HW} \rightarrow \mathbb{R}^{1 \times HW}$ and GELU to gather and reallocate channel-wise information, as shown in Equation (8):

$$\text{CA}(X) = X + \gamma_c \odot (X - \text{GELU}(XW_r)) \quad (8)$$

where γ_c is the channel-wise scaling factor initialized as zeros. It reallocates the channel-wise feature with the complementary interactions $(X - \text{GELU}(XW_r))$.

The CAEDB module is designed for $2\times$ spatial reduction. It accepts an input feature map of size $C_{in} \times H \times W$ and produces an output of size $C_{out} \times (H/2) \times (W/2)$, effectively downsampling the spatial resolution while expanding the channel count from C_{in} to C_{out} and preserving critical information.

3.3. Regional Context Feature Pyramid Network

To address the challenges of inefficient multi-scale feature fusion, complex background interference, and small-object degradation commonly encountered in underwater imagery, we propose RCFPN. This framework enhances the baseline Feature Pyramid Network (FPN) by incorporating two novel components: a SKSA mechanism and a CARAFE module, as illustrated in Figure 4.

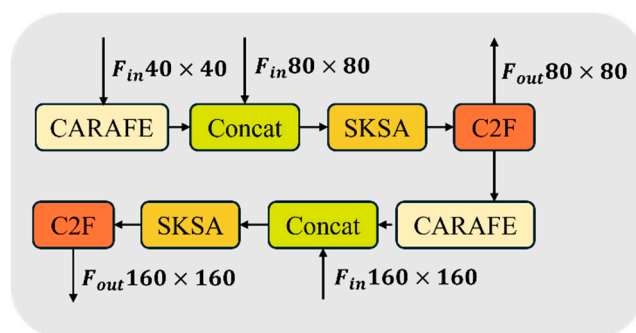


Figure 4. Regional Context Feature Pyramid Network (RCFPN) Structure diagram.

The standard FPN aggregates multi-level features primarily through direct concatenation, which often fails to emphasize salient information while introducing redundant background noise. This limitation undermines the discriminative power of subsequent decoding stages. To overcome this, the SKSA module is introduced. In contrast to conventional channel-only attention mechanisms, SKSA employs a separable kernel structure that efficiently models spatial dependencies. By adaptively emphasizing informative regions and suppressing irrelevant background responses, SKSA enhances the network's capacity to capture fine-grained details in cluttered underwater environments substantially, leading to notable improvements in detection accuracy, particularly for small and partially occluded objects.

SKSA achieves significant computational efficiency by decomposing two-dimensional convolutional kernels into two cascaded one-dimensional separable kernels. This decomposition strategy effectively reduces computational complexity while preserving the benefits of a large receptive field. Specifically, SKSA replaces standard 2D convolution with a separable structure that processes horizontal and vertical spatial dimensions independently through sequential 1D convolutions.

The input feature map F undergoes a sequential transformation through Equations (9)–(12), which collectively implement the SKSA mechanism. Specifically, Equation (9) performs depthwise separable convolution to capture spatial information from different directions, Equation (10) further expands the convolution based on fused feature Z to extract finer features and expand the receptive field, Equation (11) adjusts the number of channels to map the features into spatial attention maps, and Equation (12) achieves feature enhancement to highlight important features. This cascaded processing culminates in the output of the SKSA module, denoted as:

$$Z' = \text{Conv}_{(2d-1) \times 1} \left(\text{Conv}_{1 \times (2d-1)}(F) \right) \quad (9)$$

$$Z = \text{Conv}_{\frac{k}{d} \times 1} \left(\text{Conv}_{1 \times \frac{k}{d}}(Z') \right) \quad (10)$$

$$A = \text{Conv}_{1 \times 1}(Z) \quad (11)$$

$$\bar{F} = A \odot F \quad (12)$$

where d represents the dilation factor. Finally, the output of SKSA, F , is obtained by performing a Schur product between the input feature F and the attention feature map A , enabling dynamic enhancement of the target region and effective suppression of irrelevant regions.

Conventional upsampling approaches, such as bilinear or nearest-neighbor interpolation, utilize fixed, content-agnostic kernels that, while computationally efficient, lack adaptability to local content variations. Consequently, these methods struggle to reconstruct subtle details such as edges, textures, and small-object features that are essential for precise detection in underwater conditions. To address this issue, RCFPN integrates the CARAFE module. The core innovation of CARAFE lies in its content-adaptive upsampling strategy, which dynamically generates reassembly kernels conditioned on the local context of input features. This enables accurate reconstruction of blurred object boundaries and enhances detail preservation, crucial for robust detection in visually degraded underwater scenes.

CARAFE consists of two principal stages: predictive kernel generation and content-aware reassembly. In the first stage, it predicts an upsampling kernel tailored to each local region based on the input feature content. In the second stage, the predicted kernel is used to weight and recombine local feature patches, producing upsampled, higher-resolution outputs (i.e., feature maps with increased spatial dimensions, typically by an upsampling factor $\sigma = 2$) that retain fine structural details. The kernel prediction process is mathematically expressed in Equations (13) and (14).

$$W_{l'} = \psi(N(X_l, k_{\text{encoder}})) \quad (13)$$

$$X'_{l'} = \phi(N(X_l, k_{\text{up}}), W_{l'}) \quad (14)$$

The kernel prediction process begins by compressing the channel depth of the input features from C to C_m via a 1×1 convolution. Subsequently, a convolutional layer with a kernel size of $k_{\text{encoder}} \times k_{\text{encoder}}$ traverses the compressed feature map, producing a feature vector of length k_{up}^2 for every spatial location. Finally, this vector is reshaped into a $k_{\text{up}} \times k_{\text{up}}$ matrix and normalized using the Softmax function, yielding the content-aware reassembly kernel $W_{l'}$ specific to the target location l' .

Given a feature map X of size $C \times H \times W$ and an upsampling factor σ , CARAFE generates a new feature map X' with a size of $C \times \sigma H \times \sigma W$. For each target region $l' = (i', j')$ in the output X' , there is a corresponding region $l = (i, j)$ in the original feature map. Here, $i = i'/\sigma$, $j = j'/\sigma$. In X , the $k \times k$ neighborhood centered at l is denoted by $N(X_l, k)$. Since each source location in the input feature map X corresponds to σ^2 target positions in X' , each target position requires a reconstruction kernel of size $k_{\text{up}} \times k_{\text{up}}$ to perform the upsampling operation. This kernel helps to assemble the feature map by taking into account local contextual information around each source location.

Upon generating the reassembly kernel $W_{l'}$, CARAFE identifies a local region $N(X_l, k_{\text{up}})$ of size $k_{\text{up}} \times k_{\text{up}}$ centered at the source location l in the input feature map X . The final value at the upsampled output location is then computed as the weighted sum of all feature values within this local neighborhood, expressed as Equation (15):

$$X'_{l'} = \sum_{r=-\lfloor \frac{k_{\text{up}}}{2} \rfloor}^{\lfloor \frac{k_{\text{up}}}{2} \rfloor} \sum_{c=-\lfloor \frac{k_{\text{up}}}{2} \rfloor}^{\lfloor \frac{k_{\text{up}}}{2} \rfloor} W_{l'}(r, c) \cdot X(l + (r, c)) \quad (15)$$

Unlike conventional interpolation-based approaches relying on fixed kernels, CARAFE dynamically generates content-aware kernels (W_l') for feature reconstruction. This adaptive process not only increases the spatial resolution of feature maps but also reinforces semantic consistency, suppresses background noise, and recovers intricate details more effectively—demonstrating clear advantages in the restoration of degraded underwater imagery.

The RCFPN takes multiple feature maps from different levels of the backbone (e.g., with varying spatial sizes $H_i \times W_i$) as inputs. Through its unique fusion and up-sampling mechanisms, it outputs a set of multi-scale feature maps imbued with strong semantics and rich spatial details, which are then fed into the detection head for prediction.

4. Experiment and Results

4.1. Experimental Setup

In deep learning, variations in hardware and software configurations can significantly impact experimental outcomes, leading to inconsistencies in model performance. To ensure reproducibility and clarity, we provide detailed hardware and software specifications in Table 2.

Table 2. Computer parameters.

Configuration	Parameter
Operating system	Ubuntu 24.04
Accelerated environment	Cuda 12.8
Language	Python 3.8
Framework	Pytorch 2.2.2
GPU	RTX3060Ti (12 GB)

The training procedure utilized the following parameters: a learning rate of 0.01, a batch size of 16, and input image dimensions of 640×640 pixels after resizing. The model was trained for 200 epochs, using the “auto” optimizer, a random seed of 0, and a close_mosaic parameter set to 10.

We provide a comprehensive and detailed account of the training configuration used for all models in this study. To guarantee a fair and consistent comparison across all baseline methods and our proposed UW-YOLO-Bio, we adhered to a unified training recipe. This protocol was meticulously designed to align with common practices in object detection literature while ensuring optimal performance on underwater datasets. The configuration encompasses all critical hyperparameters, including the number of training epochs, input image dimensions, data augmentation strategies, optimizer and learning rate scheduler settings, and post-processing parameters such as Non-Maximum Suppression (NMS) IoU thresholds. By standardizing these conditions, we eliminate potential confounding factors and reinforce the validity of our comparative results. The exhaustive list of parameters is summarized in Table 3, which serves as a reference for researchers seeking to replicate our experiments or apply similar settings to related tasks.

Table 3. Unified training configuration for all comparative experiments.

Parameter	Value	Notes
Epoch	200	
Input Size	640×640	After resizing
Augmentation	Mosaic (dose = 10), Random Flip, HSV Hue/Saturation/Value/adjustment	Standard YOLO augmentation
Optimizer	SGD	Momentum = 0.9 Weight Decay = 0.0005
Learning Rate Scheduler	Linear decay from 0.01 to final LR = 0.01×0.01	
NMS IoU Threshold	0.7	
Validation IoU Threshold	0.5	For mAP calculation

4.2. Comprehensive Evaluation Framework

To evaluate the performance of our underwater object detection model rigorously, we employ a comprehensive set of metrics that assess both detection accuracy and computational efficiency. These metrics correspond to Equations (16)–(20), respectively. These metrics are widely used in computer vision research and provide a well-rounded view of model performance, which is especially critical for assessing robustness in challenging underwater environments.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (16)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (17)$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (18)$$

$$mAP = \frac{\sum_{i=1}^C AP_i}{C} \quad (19)$$

$$AP = \int_0^1 P(R) dR = \sum_{i=0}^{n-1} P_i \Delta R(i) = \frac{\sum P_i}{r_i} \quad (20)$$

Here, TP (True Positives) refers to the correctly detected instances, FP (False Positives) represents instances where the negative class is misclassified as positive, FN (False Negatives) indicates instances where the positive class is incorrectly identified as negative. We use Precision to measure the model ability to avoid false positives, and Recall to assess its capability in detecting all relevant objects. The F_1 -score combines both precision and recall into a single metric, offering a balanced measure of classification performance. Each class has its own precision and recall values, and plotting them on a precision-recall (P-R) curve helps visualize the trade-off between precision and recall. The area under the P-R curve is termed AP (Average Precision), which reflects the precision of the detection algorithm. Finally, we compute mAP (mean Average Precision) by averaging the AP scores across all classes.

4.3. Dataset Introduction

We evaluated our method on three challenging underwater datasets: DUO [44], RUOD [45], and URPC [46]. These datasets encompass diverse underwater scenarios, including haze effects, color distortion, light interference, and complex background conditions, providing a comprehensive test platform for our method.

DUO (Underwater Object Detection Dataset): This available dataset is specifically designed for underwater object detection tasks publicly. It contains 7782 images with 74,515 annotated instances across four marine species. The dataset is divided into a training set (5447 images), a validation set (778 images), and a test set (1557 images). DUO offers multi-resolution images (ranging from 3840×2160 to 586×480 pixels) to accommodate varying computational resource constraints. It is particularly valuable for testing

challenges such as low contrast, small object scales, and background clutter in underwater environments. DUO is widely used for real-time visual systems in autonomous underwater vehicles (AUVs) and serves as a critical benchmark for evaluating model generalization and multi-scale feature learning capabilities. DUO dataset is available at <https://github.com/chongweiliu/DUO> (accessed on 1 October 2025).

RUOD (Robust Underwater Object Detection Dataset): RUOD is a large-scale dataset that includes 14,000 high-resolution images with 74,903 annotated instances across 10 categories. The dataset is divided into a training set (9800 images), a validation set (2100 images), and a test set (2100 images). It simulates real-world conditions such as fog effects, color shifts, and illumination variations, thus testing model robustness under challenging noise, blur, and color distortion. RUOD is a prominent benchmark in the field, emphasizing detection performance across multi-scale variations, occlusions, and complex backgrounds, and is widely used to evaluate advanced deep learning models for applications in resource exploration and environmental monitoring. RUOD dataset is available at <https://github.com/dlut-dimt/RUOD> (accessed on 1 October 2025).

URPC (Underwater Robot Perception Challenge): This dataset is derived from real-world scenarios at Zhangzidao Marine Ranch in Dalian, China, and includes 5543 images annotated with four economically significant marine species. The dataset is divided into a training set (3880 images), a validation set (554 images), and a test set (1109 images). URPC focuses on applications in intelligent fishery and automated harvesting tasks. The dataset is notable for its practical applicability, as the images were captured in real operational environments, which include challenges such as particulate interference and object overlap. It provides valuable resources for training and testing the visual systems of underwater robots, contributing to improved efficiency in fishery resource management and automation. URPC dataset is available at <https://github.com/mousecpn/DG-YOLO> (accessed on 1 October 2025).

4.4. Ablation Study

A comprehensive set of ablation experiments was conducted on the DUO, RUOD, and URPC datasets to evaluate the contributions of the three key modules systematically: the GCPM, the CAEDB, and the RCFPN. The experiments aimed to quantify the impact of each module on critical performance metrics, including Precision, Recall, F1-Score, mAP50, and mAP50-95. For consistency, all models were trained under identical conditions, adhering to the hyperparameters and hardware specifications outlined in Section 4.1. The evaluation protocol started with a baseline model using YOLOv8n, followed by incremental additions of the proposed modules. The results, summarized in Table 4, show the average performance over multiple independent runs to ensure statistical reliability.

Impact of GCPM: The GCPM consistently improved Precision and mAP50 across all datasets, with the most notable increase observed on RUOD (a 1.2% improvement in mAP50). By integrating global context modeling and 3D perceptual enhancement, this module effectively mitigated environmental degradations such as haze and light interference. However, it occasionally led to slight reductions in Recall (e.g., a − 0.1% decrease on DUO), as it tends to prioritize semantic features over finer, localized details.

Impact of CAEDB: The CAEDB module notably enhanced Recall and F1-score, especially on RUOD (a 2.0% improvement in Recall). Through efficient channel aggregation during downsampling, it preserved critical features and improved information flow. The module excelled under noisy and occluded conditions, as observed in the challenging environmental subsets of RUOD.

Table 4. Ablation study results on DUO, RUOD, and URPC datasets.

Dataset	Configuration	Precision (%)	Recall (%)	F1-Score (%)	mAP50 (%)	mAP50-95 (%)
DUO	Baseline	84.5	82.1	83.3	85.4	50.1
DUO	+GCPM	85.6	82.0	83.8	86.4	50.8
DUO	+CAEDB	85.0	83.5	84.2	87.2	51.2
DUO	+RCFPN	86.1	83.8	84.9	88.1	51.9
DUO	Full Model	86.7	84.3	85.5	88.0	52.3
RUOD	Baseline	80.0	74.1	76.9	80.5	45.5
RUOD	+GCPM	81.7	73.9	77.7	81.7	46.2
RUOD	+CAEDB	80.9	76.1	78.4	82.2	47.2
RUOD	+RCFPN	82.5	76.3	79.3	83.8	48.1
RUOD	Full Model	82.5	76.5	79.4	83.3	48.6
URPC	Baseline	81.0	75.0	77.9	82.0	47.0
URPC	+GCPM	82.0	75.5	78.7	83.0	47.8
URPC	+CAEDB	81.5	76.5	78.9	83.5	48.2
URPC	+RCFPN	82.8	76.8	79.7	84.8	49.0
URPC	Full Model	83.0	77.0	79.9	85.0	49.5

Impact of RCFPN: The RCFPN module delivered the most significant improvements in both mAP50 and mAP50-95, with the largest gain on RUOD (a 3.3% increase in mAP50). By integrating SKSA and CARAFE, the RCFPN strengthened multi-scale feature fusion and upsampling, proving particularly effective for detecting small and blurred targets.

It is noteworthy that on the DUO dataset, the full model's mAP50 (88.0%) is marginally lower than that of the model with only RCFPN (88.1%). This slight discrepancy can be attributed to the inherent trade-offs between different module functionalities and statistical variance across runs. While the RCFPN alone excels in multi-scale fusion on DUO, the full model integrated approach provides more consistent and robust performance gains across all three diverse underwater datasets (DUO, RUOD, URPC), as reflected in the average improvement.

Synergistic Effect of the Full Model: The complete model, integrating all three modules, achieved the best overall performance. The GCPM and CAEDB modules provided enhanced feature representations for the RCFPN, enabling more effective feature fusion and outperforming the baseline across all metrics. For example, on URPC, mAP50 improved by 3.0%, demonstrating the complementary nature of the modules: GCPM contributes global context, CAEDB optimizes feature extraction, and RCFPN refines feature fusion. This synergy collectively addresses key challenges in underwater object detection. Additionally, the full model reduced the parameter count by 8.3% compared to the baseline, confirming enhanced efficiency.

4.5. Comparative Experiments

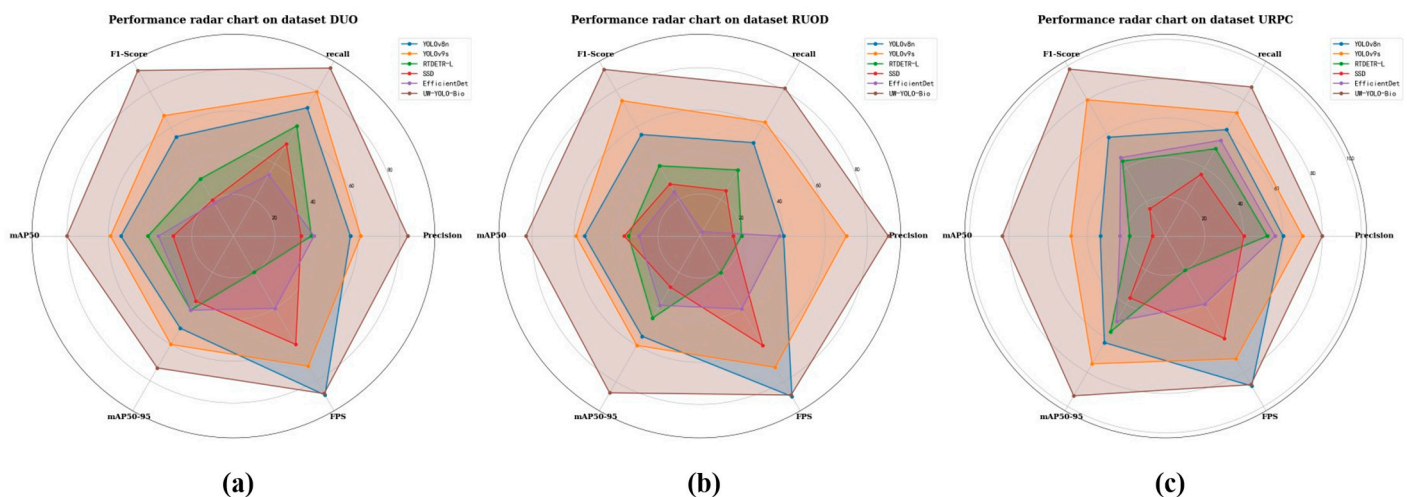
In this section, we present a comprehensive comparative evaluation of UW-YOLO-Bio against several state-of-the-art object detectors across the DUO, RUOD, and URPC datasets. The evaluated competitors include prominent detectors from the YOLO family (e.g., YOLOv8n [3], YOLOv9s [47]), as well as RT-DETR-L [3], SSD [48], and EfficientDet [20]. All models were trained and evaluated under identical conditions to ensure a fair comparison. Performance was assessed using key metrics—Precision, Recall, mAP50, and mAP50-95—with visual analyses to provide a comprehensive view of the model capabilities.

As summarized quantitatively in Table 5 and visually in the radar chart in Figure 5, the proposed method outperforms the majority of baseline models across all benchmarks, particularly excelling in mAP50 and F1-Score, surpassing the best competitor by 2.0%.

Table 5. Performance comparison between the proposed model and classification methods on datasets DUO, RUOD, and URPC.

Dataset	Method	Precision (%)	Recall (%)	F1-Score (%)	mAP50 (%)	mAP50-95 (%)	Params (M)	FPS
DUO	YOLOv8n	84.5	82.1	83.3	85.4	50.1	3.01	62.0
DUO	YOLOv9s	84.9	83	84	85.9	51	4.50	58.0
DUO	RTDETR-L	83	81.1	81.9	84.1	49.1	32.0	45.0
DUO	SSD	82.6	80.1	81.2	82.9	48.6	24.5	55.0
DUO	EfficientDet	83.1	78.4	81.1	83.6	49.1	5.30	50.0
DUO	Our Model	86.7	84.3	85.5	88.0	52.3	2.76	61.8
RUOD	YOLOv8n	80.0	74.1	76.9	80.5	45.5	3.01	62.0
RUOD	YOLOv9s	81.5	75	78.2	80.9	46	4.50	58.0
RUOD	RTDETR-L	79	72.9	75.7	78.4	44.5	32.0	45.0
RUOD	SSD	78.8	72	75	78.6	42.8	24.5	55.0
RUOD	EfficientDet	79.9	70.2	74.7	77.9	43.8	5.30	50.0
RUOD	Our Model	82.5	76.5	79.4	83.3	48.6	2.76	61.8
URPC	YOLOv8n	81.0	75.0	77.9	82.0	47.0	3.01	62.0
URPC	YOLOv9s	82	75.8	79	82.9	48	4.50	58.0
URPC	RTDETR-L	80.2	74.1	77.2	81.1	46.5	32.0	45.0
URPC	SSD	79	72.9	75.8	80.4	44.9	24.5	55.0
URPC	EfficientDet	80.6	74.5	77.3	81.4	46	5.30	50.0
URPC	Our Model	83.0	77.0	79.9	85.0	49.5	2.76	61.8

Bold markings indicate optimal performance.

**Figure 5.** Radar plots of performance of different models, (a–c) represent the performance on datasets DUO, RUOD, and URPC, respectively.

These results confirm the model robust ability to address key underwater challenges, including low contrast, noise, and significant scale variations. In terms of efficiency, the proposed model contains only 2.76 M parameters, an 8.3% reduction compared to the YOLOv8n baseline (3.01 M). Critically, as quantified in Table 5, the proposed model achieves an inference speed of 61.8 FPS while maintaining the highest accuracy. This speed significantly surpasses comparative models like RT-DETR-L (45.0 FPS) and EfficientDet (50.0 FPS), and comfortably meets the standard criterion for real-time application (typically considered as >30 FPS). The end-to-end latency breakdown for UW-YOLO-Bio: Backbone (12.3 ms), Neck with RCFPN (2.1 ms), and Head (1.8 ms), resulting in a total latency of 16.2 ms. The balance between accuracy and efficiency is driven by the lightweight design of the GCPM and CAEDB modules, underscoring the model suitability for deployment on resource-constrained underwater platforms.

4.6. Confusion Matrix

The confusion matrix offers a detailed visualization of the model classification performance, highlighting both correct predictions and misclassifications across categories. Normalized confusion matrices were generated for each dataset to evaluate the model's behavior under conditions (see Figure 6).

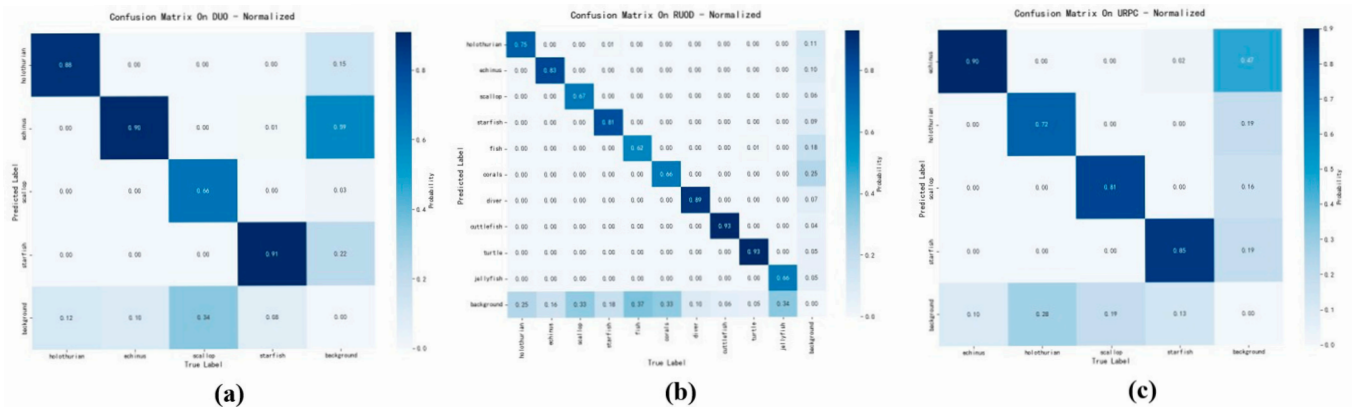


Figure 6. Confusion matrix on different datasets, (a–c) represent the performance on datasets DUO, RUOD, and URPC, respectively.

On the DUO dataset, the proposed model demonstrated high accuracy across all four categories (echinus, holothurian, starfish, scallop). The confusion matrix revealed a low false detection rate, indicating the model's effectiveness in distinguishing visually similar entities, such as holothurians and the background regions.

For the more challenging RUOD dataset, which includes ten categories, the model exhibited robust performance. It notably reduced confusion among difficult-to-detect classes such as jellyfish and corals. While fish and divers were recognized with high precision, jellyfish, due to their morphological variability and scarcity, were occasionally misclassified as background.

In the URPC dataset, derived from a marine ranch scenario, the model demonstrated balanced and reliable performance. Both precision and recall exceeded 80% for each category, reflecting its strong generalization capability, particularly in practical applications.

4.7. Results Presentation

To demonstrate the model performance in real underwater scenarios, we provide visual examples of both successful detection cases and challenging scenarios, including low contrast, occlusion, and noise.

As shown in Figure 7, visualization results show that the proposed model excels at detecting scales, even when they are mixed with sediment. On the DUO dataset, the model accurately detects small targets and overlapping objects, such as sea urchins and sea cucumbers, in low contrast images. In contrast, baseline models like YOLOv8n may miss these detections. On the RUOD dataset, the model effectively handles challenging conditions like fog effects and light interference, while reducing false positives. For example, in fog-affected images, the model successfully detects scallop, which other models may incorrectly identify as background. On the URPC dataset, the model performs consistently well in marine ranch scenes, where it detects multi-scale and partially occluded targets.

To provide a more intuitive and direct comparison of the detection performance, Figure 8 presents a comparative analysis of the baseline YOLOv8n model and our proposed UW-YOLO-Bio framework on identical challenging images selected from the test sets. Figure 8 is organized into three columns: the first column displays the ground truth annotations, the second column shows the detection results obtained by the original

YOLOv8n model, and the third column presents the results from our enhanced UW-YOLO-Bio model. Each line corresponds to a specific failure mode: (a) low confidence detections, (b) missed detections, and (c) false detections, which are common challenges in underwater object detection.

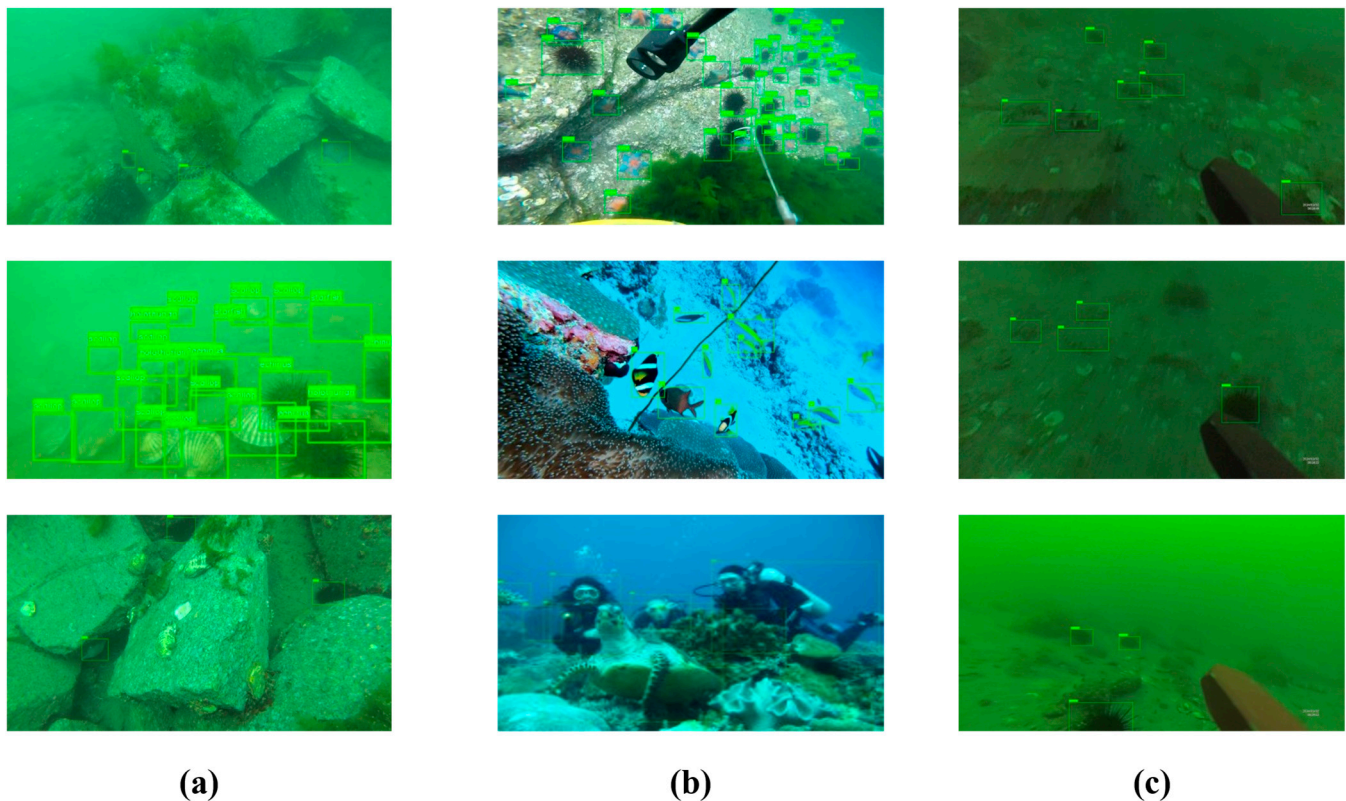


Figure 7. The detection results of the UW-YOLO-Bio. (a) For the detection results on the DUO dataset, (b) for the results on RUOD, and (c) for the results on URPC.

As illustrated in Figure 8a, which focuses on low confidence scenarios, our model significantly increases the confidence levels for various target detections, demonstrating enhanced reliability and superiority. The low confidence observed in the original YOLOv8n results can be attributed to the targets blending with the background, particularly due to shadows cast by underwater rocks that closely match the colors of both the rocks and the targets, making identification challenging. This improvement underscores the effectiveness of our global context modeling in mitigating environmental ambiguities.

In Figure 8b, which addresses missed detection cases, our model detects a greater number of correct targets, reducing omission errors. This improvement is notable given that the targets are often obscured by underwater sediment and aquatic vegetation, which poses a significant challenge for efficiently extracting key features for identification. Through a series of optimization techniques integrated into our framework, including RCFPN, we have enhanced the model's capability to handle such occlusions and complex backgrounds, thereby improving recall.

Figure 8c highlights instances where our model reduces false detection rates, particularly false positives. These false detections in the baseline model occur due to the bluish-green tint of the water, which causes objects like starfish to closely resemble the background. Our model exhibits improvements in mitigating this issue by better distinguishing between foreground objects and the ambient environment through the CAEDB and attention mechanisms, thereby reducing misclassifications and enhancing overall detection accuracy.

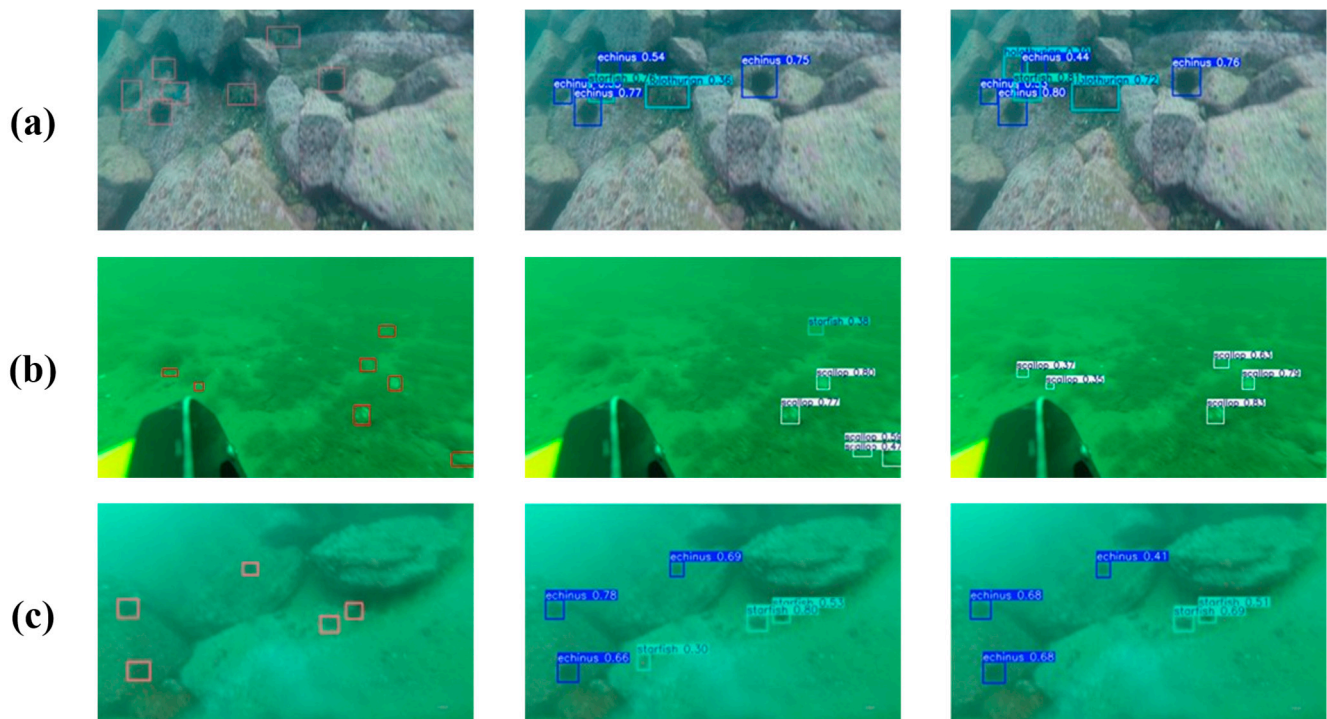


Figure 8. Comparison of detection performance with baseline model (yolov8n) under challenging underwater conditions. (a) Low confidence; (b) missed detection cases; (c) false detection.

Despite the overall excellent performance, we also analyzed the failure cases to identify areas for further improvement. As illustrated in Figure 9, these failures primarily occurred in scenarios involving extreme occlusion, high imbalance of image boundary targets, or problematic class distributions.

Firstly, as illustrated in Figure 9a, the model struggles to delineate individual contours when multiple targets are densely clustered or partially obscured by sediments, often resulting in missed detections or the fusion of multiple instances into a single bounding box. This limitation stems from the model reliance on local appearance features, which are significantly diminished under extreme occlusion, making it difficult to obtain sufficient information for effective instance separation and recognition.

Furthermore, another recurrent issue is observed in the periphery of images. Figure 9b,c demonstrate that for objects truncated by the image frame, the model detection confidence plummets, frequently leading to complete misses. The root cause for this boundary effect is a training-inference discrepancy, the proportion of complete targets in the training data is much higher than that of truncated targets, resulting in weaker generalization ability of the model to incomplete targets.

Lastly, despite the designed enhancements for low-contrast scenarios, the model remains vulnerable to highly camouflaged targets. As shown in Figure 9d, when the color and texture of a target are nearly indistinguishable from the surrounding environment, the model frequently misclassifies it as background. This represents an extreme manifestation of the low-contrast challenge, where the discriminative power of low-level features is insufficient, and the model lacks the necessary high-level semantic or contextual cues to resolve the ambiguity based on a single static frame. Collectively, these cases provide crucial insights for directing future research toward enhancing contextual reasoning and robustness in complex underwater environments.

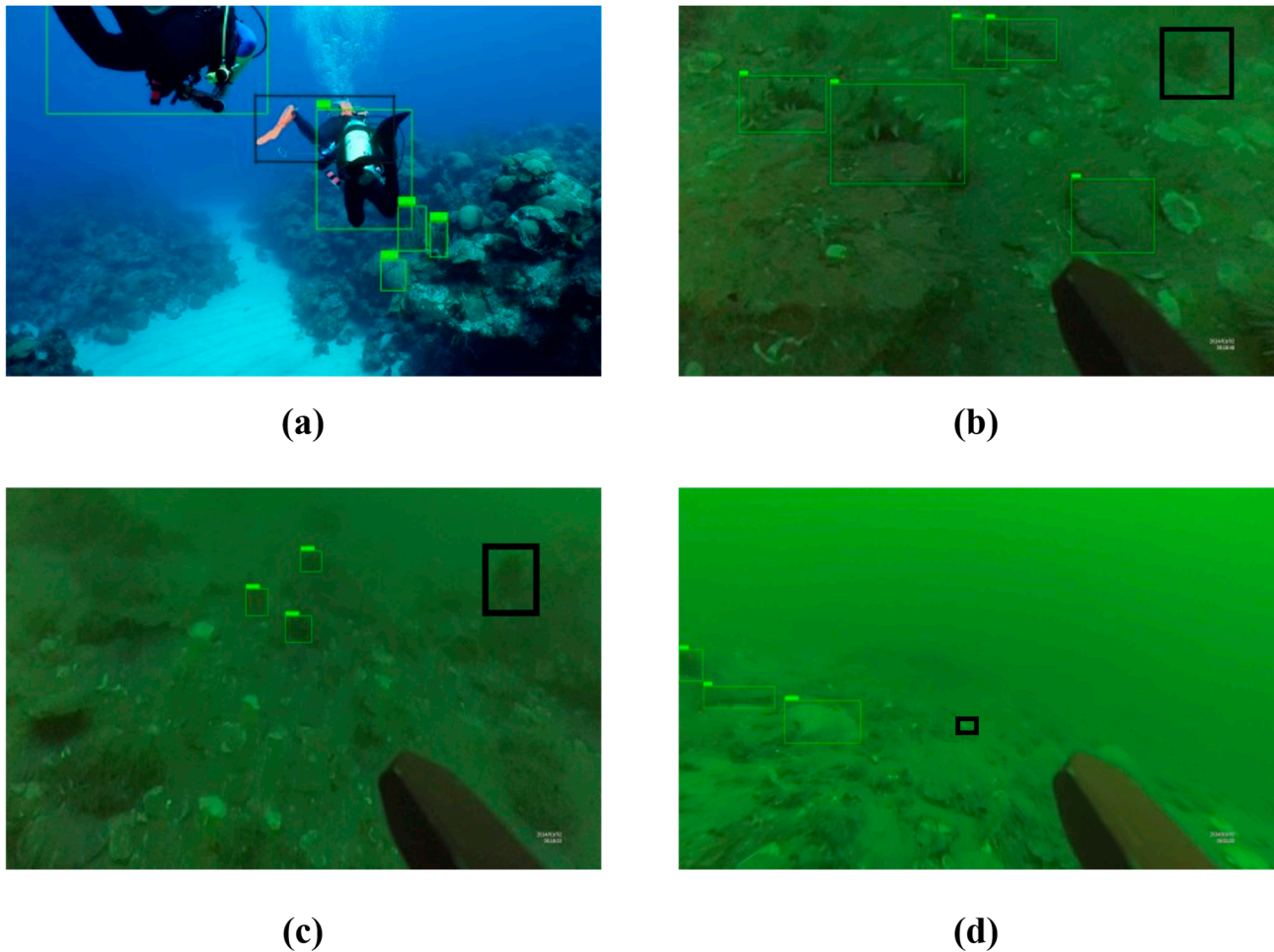


Figure 9. Failure cases. (a) is extreme occlusion, (b) and (c) are boundary blurring, and (d) is blending into the background. The black square is the target of missed detection.

Although the primary evaluation was conducted on datasets encompassing common underwater degradations (e.g., haze, blur, color cast), the model’s design inherently addresses challenges analogous to adverse weather conditions. The GCPM global context modeling enhances robustness to haze-like effects caused by high concentrations of suspended particles, similar to fog. Concurrently, the CAEDB and RCFPN focus on preserving details and suppressing noise helps maintain performance under varying illumination conditions, such as those caused by sunlight penetration at different depths or water turbidity levels. We therefore expect the model to demonstrate consistent performance across a range of water clarity and ambient light conditions typical of different ‘weather’ underwater. Explicit validation on datasets explicitly tagged with meteorological or water quality metadata remains a valuable avenue for future research.

A limitation of this work is that it does not evaluate against newer detector versions like YOLOv10. Architectural differences across generations could influence performance, and exploring the integration of our proposed modules with such advanced backbones is a valuable direction for future work.

Future Work: Based on the analysis of failure cases, several promising directions for improvement have been identified:

- **Sensitivity to Edge and Occluded Targets:** Enhancing the model sensitivity to edge and occluded targets could be achieved by incorporating edge-aware sampling strategies or more sophisticated data augmentation techniques tailored for such scenarios.

- **Class Imbalance:** Addressing class imbalance through techniques like resampling or focal loss adjustments could lead to improved performance, particularly for underrepresented classes.
- **Robustness to Occlusions:** Improving robustness against occlusions could involve integrating context-aware feature reasoning or employing multi-frame analysis to infer obscured parts of objects.
- **Validation Under Diverse Environmental Conditions:** Explicit validation on datasets explicitly tagged with meteorological or water quality metadata remains a valuable avenue for future research. This would systematically assess model robustness across varying weather conditions, water clarity, and illumination scenarios, enhancing generalizability for real-world deployments.
- **Integration with Advanced Detector Architectures:** Architectural differences across generations could significantly influence performance. Therefore, a valuable future direction is to explore the integration of the proposed modules with state-of-the-art backbones such as YOLOv10. This would leverage advancements in efficiency and accuracy, potentially leading to further improvements in real-time underwater detection while maintaining the lightweight design principles of our framework.

5. Conclusions

This study introduces an advanced underwater object detection framework that integrates three novel components into the YOLOv8 architecture: GCPM, CAEDB, and RCFPN. Evaluated across three challenging underwater datasets—DUO, RUOD, and URPC—the proposed method achieves state-of-the-art performance, with significant gains in both accuracy and computational efficiency. Key results include: mAP50 scores of 88.0% on DUO, 83.3% on RUOD, and 85.0% on URPC. A reduction of 8.3% in the number of parameters, making the model more efficient. Real-time inference at 61.8 FPS, enabling practical deployment in underwater environments. The model demonstrates strong robustness against common underwater challenges, such as low contrast, noise, and occlusion, making it a highly suitable candidate for deployment on autonomous underwater vehicles (AUVs).

Author Contributions: Conceptualization, W.Z., J.Z. and S.L.; methodology, W.Z. and J.Z.; software, W.Z.; validation, W.Z.; formal analysis, S.L.; investigation, Y.Z.; resources, J.Z.; data curation, W.Z.; writing—original draft preparation, W.Z.; writing—review and editing, J.Z.; visualization, W.Z.; supervision, J.Z.; project administration, J.Z.; funding acquisition, J.Z. and S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Key Research and Development Program of China [Grant Number 2021YFC2801100] and the National Natural Science Foundation of China (NSFC) [Grant Number 42176194].

Data Availability Statement: Data will be made available on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Jocher, G.; Chaurasia, A.; Qiu, J. YOLOv8. 2023. Available online: <https://docs.ultralytics.com/zh/models/yolov8/> (accessed on 1 October 2025).
2. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. *End-to-End Object Detection with Transformers*; Springer: Cham, Switzerland, 2020; pp. 213–229.
3. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-time Object Detection. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974.

4. Jaffe, J.S. Computer Modeling and the Design of Optimal Underwater Imaging Systems. *IEEE J. Ocean. Eng.* **1990**, *15*, 101–111. [[CrossRef](#)]
5. Li, J.; Skinner, K.A.; Eustice, R.M.; Johnson-Roberson, M. WaterGAN: Unsupervised Generative Network to Enable Real-Time Color Correction of Monocular Underwater Images. *IEEE Robot. Autom. Lett.* **2018**, *3*, 387–394. [[CrossRef](#)]
6. Hu, Y.; Wang, Z.; Zhang, Q.; Huang, W.; Zheng, B. UWaterDetect: A Lightweight Underwater Object Detection Model based on RT-DETR and Wavelet Convolutions. In Proceedings of the OCEANS 2025 Brest, Brest, France, 16–19 June 2025; pp. 1–6.
7. Islam, M.J.; Xia, Y.; Sattar, J. Fast Underwater Image Enhancement for Improved Visual Perception. *IEEE Robot. Autom. Lett.* **2020**, *5*, 3227–3234. [[CrossRef](#)]
8. Zhang, W.; Wang, H.; Ren, P.; Zhang, W. Underwater Image Color Correction via Color Channel Transfer. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1500805. [[CrossRef](#)]
9. Feng, J.; Jin, T. LMFEN: Lightweight multi-scale feature enhancement network for underwater object detection in AUVs. *Ocean. Eng.* **2025**, *339*, 122109. [[CrossRef](#)]
10. Shang, H.; Yu, Y.; Song, W. Underwater fish image classification algorithm based on improved ResNet-RS model. In Proceedings of the Jiangsu Annual Conference on Automation (JACA 2023), Changzhou, China, 10–12 November 2023; pp. 106–110.
11. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. *CBAM: Convolutional Block Attention Module*; Springer: Cham, Switzerland, 2018; pp. 3–19.
12. Liu, K.; Peng, L.; Tang, S. Underwater Object Detection Using TC-YOLO with Attention Mechanisms. *Sensors* **2023**, *23*, 2567. [[CrossRef](#)] [[PubMed](#)]
13. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
14. Xu, X.; Jiang, Y.; Chen, W.; Huang, Y.; Zhang, Y.; Sun, X. DAMO-YOLO: A Report on Real-Time Object Detection Design. *arXiv* **2022**, arXiv:2211.15444. [[CrossRef](#)]
15. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 9992–10002.
16. Mehta, S.; Rastegari, M.J.A. MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer. *arXiv* **2021**, arXiv:2110.02178.
17. Li, Y.; Yuan, G.; Wen, Y.; Hu, J.; Evangelidis, G.; Tulyakov, S.; Wang, Y.; Ren, J. EfficientFormer: Vision Transformers at MobileNet Speed. *arXiv* **2022**, arXiv:2206.01191. [[CrossRef](#)]
18. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
19. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
20. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787.
21. Li, C.; Anwar, S.; Porikli, F. Underwater scene prior inspired deep underwater image and video enhancement. *Pattern Recognit.* **2020**, *98*, 107038. [[CrossRef](#)]
22. Golubović, D.; Erić, M.; Vukmirović, N. High-Resolution Azimuth Detection of Targets in HFSWRs Under Strong Interference. In Proceedings of the 2024 11th International Conference on Electrical, Electronic and Computing Engineering (IcETRAN), Nis, Serbia, 3–6 June 2024; pp. 1–6.
23. Golubović, D.; Erić, M.; Vukmirović, N. High-Resolution Doppler and Azimuth Estimation and Target Detection in HFSWR: Experimental Study. *Sensors* **2022**, *22*, 3558. [[CrossRef](#)]
24. Golubović, D.; Vukmirović, N.; Erić, M. An Introduction to Vessel Tracking in HFSWRs Based on a High-Resolution Range-Doppler Map: Some Preliminary Results and Challenges. In Proceedings of the 2024 13th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro, 11–14 June 2024; pp. 1–5.
25. Ji, X.; Liu, G.-P.; Cai, C.-T. Collaborative Framework for Underwater Object Detection via Joint Image Enhancement and Super-Resolution. *J. Mar. Sci. Eng.* **2023**, *11*, 1733. [[CrossRef](#)]
26. Du, D.; Si, L.; Xu, F.; Niu, J.; Sun, F. A Physical Model-Guided Framework for Underwater Image Enhancement and Depth Estimation. *arXiv* **2024**, arXiv:2407.04230. [[CrossRef](#)]
27. Zhuang, P.; Li, C.; Wu, J. Bayesian retinex underwater image enhancement. *Eng. Appl. Artif. Intell.* **2021**, *101*, 104171. [[CrossRef](#)]
28. Dai, L.; Liu, H.; Song, P.; Liu, M. A gated cross-domain collaborative network for underwater object detection. *Pattern Recognit.* **2024**, *149*, 110222. [[CrossRef](#)]
29. Gao, X.; Yang, M.; Xie, Z. Domain Adaptive Underwater Object Detection via Complementary Style-Aware Learning. *Neural Netw.* **2025**, *194*, 108174. [[CrossRef](#)] [[PubMed](#)]

30. Wang, S.; Park, S.; Kim, J.; Kim, J. Safety Helmet Monitoring on Construction Sites Using YOLOv10 and Advanced Transformer Architectures with Surveillance and Body-Worn Cameras. *J. Constr. Eng. Manag.* **2025**, *151*, 04025186. [\[CrossRef\]](#)
31. Wang, S.J. Automated non-PPE detection on construction sites using YOLOv10 and transformer architectures for surveillance and body worn cameras with benchmark datasets. *Sci. Rep.* **2025**, *15*, 27043. [\[CrossRef\]](#)
32. Kang, M.; Ting, C.-M.; Fung Ting, F.; Phan, R. CST-YOLO: A Novel Method for Blood Cell Detection Based on Improved YOLOv7 and CNN-Swin Transformer. *arXiv* **2023**, arXiv:2306.14590. [\[CrossRef\]](#)
33. Cao, Y.; Li, C.; Peng, Y.; Ru, H. MCS-YOLO: A Multiscale Object Detection Method for Autonomous Driving Road Environment Recognition. *IEEE Access* **2023**, *11*, 22342–22354. [\[CrossRef\]](#)
34. Jiang, H.; Zhong, J.; Ma, F.; Wang, C.; Yi, R. Utilizing an Enhanced YOLOv8 Model for Fishery Detection. *Fishes* **2025**, *10*, 81. [\[CrossRef\]](#)
35. Zhang, W.; Wang, H.; Li, H.; Li, Y.; Song, T.; Liu, H.; Wang, G. Dual-stream feature pyramid network with task interaction for underwater object detection. *Digit. Signal Process.* **2025**, *163*, 105199. [\[CrossRef\]](#)
36. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [\[CrossRef\]](#)
37. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
38. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767. [\[CrossRef\]](#)
39. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861. [\[CrossRef\]](#)
40. Li, S.; Wang, Z.; Liu, Z.; Tan, C.; Lin, H.; Wu, D.; Chen, Z.; Zheng, J.; Li, S.Z. MogaNet: Multi-order Gated Aggregation Network. *arXiv* **2024**, arXiv:2211.03295.
41. Zhou, H.; Kong, M.; Yuan, H.; Pan, Y.; Wang, X.; Chen, R.; Lu, W.; Wang, R.; Yang, Q. Real-time underwater object detection technology for complex underwater environments based on deep learning. *Ecol. Inform.* **2024**, *82*, 102680. [\[CrossRef\]](#)
42. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
43. Yang, L.; Zhang, R.-Y.; Li, L.; Xie, X. Simam: A simple, parameter-free attention module for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 11863–11874.
44. Liu, C.; Li, H.; Wang, S.; Zhu, M.; Wang, D.; Fan, X.; Wang, Z. A Dataset and Benchmark of Underwater Object Detection for Robot Picking. In Proceedings of the 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Shenzhen, China, 5–9 July 2021; pp. 1–6.
45. Fu, C.; Liu, R.; Fan, X.; Chen, P.; Fu, H.; Yuan, W.; Zhu, M.; Luo, Z. Rethinking general underwater object detection: Datasets, challenges, and solutions. *Neurocomputing* **2023**, *517*, 243–256. [\[CrossRef\]](#)
46. Liu, H.; Song, P.; Ding, R. WQT and DG-YOLO: Towards domain generalization in underwater object detection. *arXiv* **2020**, arXiv:2004.06333. [\[CrossRef\]](#)
47. Shi, Y.; Duan, Z.; Qing, S.; Zhao, L.; Wang, F.; Yuwen, X. YOLOv9s-Pear: A Lightweight YOLOv9s-Based Improved Model for Young Red Pear Small-Target Recognition. *Agronomy* **2024**, *14*, 2086. [\[CrossRef\]](#)
48. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. *SSD: Single Shot MultiBox Detector*; Springer: Cham, Switzerland, 2016; pp. 21–37.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.