mathematics

MDPI

Article

YOLO-CAB: An Efficient Deep Learning-Based Underwater Object Detection Method for Autonomous Underwater Vehicles

Runze Li [1], Changdong Yu [2,*], Shuaiyu Bao [3], Zijian Li [2] and Jinyi Yao [4]

[1] College of Marine Electrical Engineering, Dalian Maritime University, Dalian 116026, China; tearslee@dlmu.edu.cn
[2] College of Artificial Intelligence, Dalian Maritime University, Dalian 116026, China; lizj@dlmu.edu.cn
[3] College of Shipping Economics and Management, Dalian Maritime University, Dalian 116026, China; baoshuaiyu2005@dlmu.edu.cn
[4] School of Economics & Management, Dalian Minzu University, Dalian 116600, China; yjy@dlnu.edu.cn
* Correspondence: ycd@dlmu.edu.cn

Abstract

High-precision environmental perception is essential for deep-sea exploration and autonomous underwater vehicle operations. However, physical factors such as light scattering and selective absorption, along with high target–background similarity, cause existing detection methods to suffer from low recall and inaccurate localization on targets with blurred edges, low contrast, or category ambiguity. To address these challenges, we propose YOLO-CAB, a YOLOv13-based underwater object detector designed to handle these complex scenes by optimizing feature extraction, boundary perception, and the loss function. First, we introduce the Context-Aware Large Selective Kernel (CALSK) module into the shallow backbone layers to expand the receptive field and adaptively enhance spatial features via multi-scale depthwise convolutions. Furthermore, the Spatial Boundary Attention Module (SBAM) is applied before the feature pyramid enters the detection head to refine multi-scale features and enhance sensitivity to target boundaries. To address the variance in detection difficulty across categories, we also develop the Momentum-based Category-Aware Weighted Intersection over Union (MCAWIoU) loss. Consequently, the proposed weighting mechanism improves localization accuracy and confidence distribution for challenging samples. Evaluated on the RUOD benchmark dataset, YOLO-CAB improves mean average precision (mAP_{50}) by 4.67%, the stricter localization metric ($mAP_{50–95}$) by 3.0%, and F_1 score by 3.7% over the vanilla YOLOv13n baseline. Ablation studies confirm the individual and synergistic contributions of these components. With 8.85 GFLOPs and an inference time of 5.21 ms under the tested hardware setting, YOLO-CAB improves detection accuracy while maintaining real-time inference.

Keywords: underwater object detection; autonomous underwater vehicle (AUV); feature extraction; boundary perception; deep learning

MSC: 68T07

Academic Editor: Renato Bruni

Received: 9 April 2026
Revised: 12 May 2026
Accepted: 26 May 2026
Published: 2 June 2026

1. Introduction

As the global blue economy agenda advances, marine resource exploration has expanded from shallow waters to deep-sea and extreme environments. Unmanned underwater vehicles (UUVs) [1,2], particularly autonomous underwater vehicles (AUVs), are

critical platforms for tasks such as docking with subsea stations, surveying seabed minerals, and conducting marine ecological surveys. The intelligence of these vehicles dictates the scope and depth of underwater operations [3,4]. During complex missions, AUVs must execute millisecond-level responses to dynamic environments, relying solely on onboard perception systems. While acoustic sensors like synthetic aperture sonar excel in long-range detection [5], optical vision remains an irreplaceable modality for close-range operations and species classification due to its high resolution, low cost, and rich color information [6,7].

Transferring advanced computer vision algorithms from controlled terrestrial environments to dynamic underwater spaces presents severe physical challenges. The classical Jaffe-McGlamery imaging model establishes that underwater visual quality is fundamentally limited by medium absorption and scattering [8]. The selective absorption of long-wavelength light (e.g., red and orange) results in a pronounced red-green color cast. Backscattering from suspended particles creates a strong veiling effect, which compresses target contrast and blurs the boundaries between biological targets and the background [9,10]. This image degradation causes traditional operators like Canny and Sobel to fail. Therefore, developing robust underwater object detection methods remains a critical challenge.

In recent years, object detection research has evolved from traditional handcrafted feature-based methods to data-driven deep representations. While early two-stage networks achieve high accuracy [11], their substantial memory footprint conflicts with the strict size, weight, and power constraints of AUVs [12]. YOLO-series algorithms mitigate this by framing detection as a single regression problem [13]. Nevertheless, existing strategies exhibit limitations in specific underwater environments. In degraded underwater scenes, generic YOLO models remain less reliable for three main reasons. First, standard convolutions in shallow layers rely on fixed local receptive fields and therefore struggle to capture sufficient context for low-contrast targets with large scale variation. Second, pyramid-based feature fusion in the neck can smooth weak contours and suppress high-frequency boundary cues, reducing localization reliability for camouflaged or blurred organisms [14]. Some studies rely on preprocessing techniques like dehazing and color restoration, which introduce computational latency and risk of distorting original features [15–17]. Although advanced deep learning architectures, such as feature aggregation networks tailored for maritime environments [18–20], exhibit strong perception capabilities for marine targets, their substantial computational and structural overhead may limit high-frame-rate real-time detection on AUV-oriented edge platforms. Alternatively, increasing model depth compromises the real-time inference speed necessary for operations on onboard platforms like NVIDIA Jetson devices. Third, the standard localization objective of one-stage detectors does not adequately address long-tail class imbalance or variation in sample difficulty, which further weakens performance on rare underwater categories [21]. These limitations motivate the design of YOLO-CAB, in which CALSK strengthens multi-scale contextual perception, SBAM refines degraded boundary information, and MCAWIoU improves optimization for difficult and underrepresented categories.

In addition to computational constraints, the robustness of underwater object detection is affected by data distribution and biological evolution. Unlike terrestrial datasets with uniform illumination and balanced categories, such as COCO and ImageNet, targets in authentic underwater environments frequently exhibit strong biological camouflage, known as crypsis. Many marine organisms blur their contours by simulating the surrounding color and texture, reducing the discriminability between targets and backgrounds in the feature space [22]. This low-contrast boundary information is highly susceptible to feature loss during successive downsampling in deep convolutional networks, causing low recall for camouflaged targets. Moreover, the sparsity of marine life causes underwater datasets to

exhibit a severe long-tail distribution, complicating balanced optimization across categories. In practical AUV operations, common fish are frequently observed, whereas benthic targets such as corals and sea cucumbers are less represented because of their localized distribution. When handling these blurred boundaries and small targets, traditional intersection over union (IoU) losses, such as CIoU and DIoU, often suffer from gradient vanishing or oscillation. This instability arises because the overlap between predicted and ground truth boxes becomes volatile under noise, limiting model generalization on difficult categories. Therefore, enhancing selective attention to degraded boundaries and dynamically balancing class weights, while maintaining a lightweight architecture, remains a critical challenge for AUV visual perception [23].

The scarcity of high-quality annotated data and limited data diversity have long restricted the deployment of underwater detection algorithms on AUVs. Early research primarily relied on elementary datasets like Fish4Knowledge (F4K) [24]. Although F4K contains millions of fish images, the data mostly originates from long-term surveillance videos captured by fixed cameras. The static backgrounds and low resolution fail to simulate the complex motion blur and perspective changes AUVs encounter during dynamic navigation. Later, datasets from the Underwater Robot Picking Contest (URPC) series [25] gained widespread use. URPC provides high-resolution annotations for benthic organisms like sea cucumbers and sea urchins, advancing underwater grasping algorithms. However, URPC collection scenes are largely limited to nearshore artificial aquaculture areas with favorable lighting. They lack representations of extremely degraded deep-water environments, creating a generalization gap for models trained in these settings when deployed in complex open waters. To better simulate authentic marine ecosystems, researchers introduced task-specific datasets. For example, the BrackishMOT dataset [26] focuses on turbid estuaries, emphasizing model robustness under low visibility. The TrashCan dataset [27] focuses on marine debris monitoring, capturing interactions between waste and biological targets. While valuable for specific domains, these datasets often suffer from limited category diversity or scene homogeneity. Consequently, they cannot comprehensively evaluate AUV perception for large-scale biodiversity surveys or complex deep-sea missions. The recently released Real-world Underwater Object Detection (RUOD) benchmark [28] provides a highly representative solution for evaluating algorithms in complex marine environments. RUOD encompasses multiple physical scales and extreme degradation scenes, reflecting the long-tail distribution and biological camouflage challenges of deep-sea operations. Despite this comprehensive benchmark, existing general object detectors struggle to achieve satisfactory accuracy on it. This bottleneck arises because standard architectures lack optimization mechanisms targeting specific physical degradations, such as edge blurring, and the severe category imbalance inherent to deep-sea data.

To address these challenges, we propose YOLO-CAB, a YOLOv13-based underwater object detector tailored for complex imaging conditions, with AUV edge deployment as a target application. In this work, YOLOv13n denotes the vanilla baseline, whereas YOLO-CAB denotes the complete detector with CALSK, SBAM, MCAWIoU, and the restructured detection head. Unlike approaches that mainly rely on image preprocessing, generic attention blocks, or deeper feature aggregation, this work focuses on how underwater degradation affects different stages of the detection pipeline. Specifically, low contrast weakens shallow contextual representation, contour blurring reduces boundary reliability before prediction, and long-tailed category distributions introduce localization imbalance during optimization. The design links these three error sources to three corresponding components: weak shallow context is handled by CALSK, blurred boundaries before prediction are refined by SBAM, and category-dependent localization imbalance is addressed by MCAWIoU.

The main contributions of this work are summarized as follows:

- A YOLOv13-based underwater detector, YOLO-CAB, is developed for AUV-oriented visual perception under optical degradation. The framework addresses low contrast, blurred contours, and imbalanced category difficulty through context-aware feature extraction, boundary-sensitive refinement, and category-aware localization optimization.
- CALSK is introduced into the shallow backbone layers. It combines multi-scale depthwise convolution, additive feature fusion, and spatial weighting to improve contextual feature representation under low contrast and scale variation.
- SBAM is designed before the detection head. It extracts local edge cues and broader contour responses through parallel branches and generates a boundary-sensitive spatial mask to refine multi-scale features.
- An MCAWIoU loss is developed by tracking category-level IoU with an exponential moving average, enabling category-aware weighted localization optimization for challenging samples and underrepresented categories. Experiments and ablation studies on the RUOD benchmark dataset evaluate the individual and combined effects of the proposed components.

2. Related Work

2.1. Underwater Object Detection Methods

Underwater object detection is a key technology for enabling intelligent operations of autonomous underwater vehicles. Prior to the widespread adoption of deep learning, early underwater machine vision primarily relied on handcrafted feature extractors. Researchers conventionally utilized low-level visual features, including color histograms, gray-level co-occurrence matrices, and the histograms of oriented gradients, in conjunction with traditional machine learning algorithms, such as support vector machines, to identify marine organisms and artificial targets [29]. However, the marine environment exhibits light absorption, medium scattering, and dynamic non-uniform illumination, which severely constrain the representational capacity of handcrafted features. Consequently, traditional methods demonstrate weak generalization capabilities, which frequently result in a high rate of missed detections and false positives when models process camouflaged benthic organisms or complex seabed backgrounds.

With the application of deep convolutional neural networks, such as residual networks (ResNet), data-driven frameworks of deep learning have transformed the technical trajectory of underwater object detection [30]. Early research predominantly relied on two-stage architectures of object detection, which are exemplified by Faster R-CNN. These networks pre-generate candidate boxes via a region proposal network (RPN) and subsequently utilize deep convolutional layers for the classification of regions and the regression of bounding boxes. Two-stage networks achieve high detection accuracy on static underwater datasets and effectively identify benthic organisms, such as sea cucumbers and sea urchins. However, the requirement to process numerous candidate regions entails a substantial number of parameters and intensive computational demands. For micro-AUVs that rely on the power of onboard batteries and nodes of edge computing, such as the NVIDIA Jetson platform, these constraints regarding size, weight, and power (SWaP) fail to support the high-frequency real-time inference required by two-stage algorithms. Consequently, these algorithms fall short of the control requirements for dynamic obstacle avoidance and autonomous operations. In addition, when unmanned underwater vehicles perform delicate grasping or docking tasks, these vehicles also need to combine visual information for the real-time estimation of high-precision poses. This combination places more stringent requirements on onboard algorithms for the computation of low latency, and on networks for the capabilities of adaptive constraints in degraded underwater environments [31].

To balance the accuracy of detection and the speed of inference, one-stage detection frameworks, such as SSD [32] and the You Only Look Once (YOLO) series, have gradually emerged as the mainstream approach in underwater visual perception [13]. These frameworks eliminate the process of region proposal by redefining the detection of objects as a direct spatial regression problem, which directly outputs the probabilities of categories and the coordinates of bounding boxes from the entire image. This network topology reduces memory-access overhead and inference latency. Recent studies have proposed various improvements to the YOLO architecture for marine and underwater environments. For instance, Yu et al. propose the YOLO-MRS model [33] for the perception of unmanned surface vehicles (USVs). This model integrates a multi-scale cross-axis attention mechanism into the backbone to capture the global dependencies of features. Furthermore, this model incorporates refocused convolutions to suppress the background noise of the marine environment while maintaining a lightweight structure, which achieves an effective trade-off between the accuracy of detection and computational efficiency. Similarly, Qin et al. proposed the YOLO8-FASG method [34], which integrates a global attention mechanism and lightweight convolutions to achieve high-accuracy fish identification.

Although YOLO architectures improve the feasibility of applications in marine robotics, their direct deployment in degraded deep-sea environments still suffers from limited adaptability. This limitation mainly arises from the fact that general YOLO models are originally designed for clear terrestrial scenes. Consequently, their standard convolutional layers and feature pyramid structures, such as FPN and PANet, lack dedicated mechanisms to preserve weak and blurred underwater boundaries, which often leads to the over-smoothing of features and the loss of high-frequency spatial information [14]. Furthermore, this structural deficiency is exacerbated by the long-tail distribution of underwater data. In such scenarios, rare species coexist with a large number of background samples, making conventional one-stage detectors prone to biased learning. As a result, the localization precision for camouflaged and infrequent targets is significantly degraded [21]. These observations indicate that improving underwater detection performance requires targeted enhancements in low-level receptive field modeling, boundary feature refinement, and loss function design. These requirements motivate the selection of YOLOv13 as the baseline network and the subsequent structural reconfiguration in this study.

2.2. Feature Enhancement and Target Perception

Autonomous underwater vehicles rely on visual systems to provide the high-precision perception and localization of targets during terminal tasks, such as benthic biological surveys, object grasping, and seabed infrastructure inspection. However, underwater vision faces specific physical interferences. The selective absorption of light by the aqueous medium and the scattering from suspended particles not only diminish the contrast of images and induce the distortion of color, but also degrade the boundary contours of targets in turbid environments. The resulting loss of high-frequency spatial signals, such as biological textures and geometric edges, causes deep-sea mimetic organisms, such as sea cucumbers and flounders, to blend seamlessly into the complex backgrounds of the seabed. This blending significantly complicates the extraction of effective discriminative features by models of deep learning [35].

To alleviate the attenuation of features caused by underwater visual degradation, recent advancements in object detection incorporate feature enhancement and mechanisms of attention. One prominent approach focuses on the static or adaptive expansion of receptive fields within the network to capture richer global contextual information in environments of low contrast. Conventional deep convolutional neural networks typically rely on local kernels of 3×3. This limited receptive field encounters difficulties in the

capture of global features when models process targets with diverse scales and incomplete local features. Consequently, several studies introduce dilated convolutions or kernels of a large size. For instance, Ding et al. [36] demonstrate in RepLKNet that kernels of 31×31 provide a global field of view comparable to Transformers, thereby enhancing the robustness of downstream vision tasks. Within the underwater domain, research also utilizes multi-branch atrous spatial pyramid pooling to aggregate multi-scale contextual semantics for the inference of positions for blurred targets [37]. Recent studies also indicate that the integration of the Transformer architecture with a dual attention mechanism for multi-scale feature fusion can effectively remove the interference of background in complex surface or underwater environments. This integration exhibits excellent robustness to noise and advantages in the global receptive field for the instance-level perception of marine targets [38].

Another category of methods focuses on the recalibration of local features, which utilizes mechanisms of attention to enhance critical features while suppressing background noise. For instance, the incorporation of Squeeze-and-Excitation (SE) modules or Convolutional Block Attention Modules (CBAM) into the backbone network allows for the evaluation of correlations between channels or spatial pixels [39,40]. This approach assigns higher weights to high-response regions, thereby suppressing noise interference caused by underwater backscattering. An improved YOLO algorithm proposed by Zhang et al. [41] integrates coordinate attention to enhance the capabilities of cross-channel localization within the network regarding the feature maps of underwater targets.

Although these strategies for feature enhancement improve the recall of detection for underwater targets to some extent, structural limitations persist under the computational constraints of micro-AUVs and the characteristics of authentic long-tail scenarios in the deep sea. One primary issue is the lack of flexibility in the mechanisms for the expansion of receptive fields. The direct stacking of dilated convolutions is prone to the gridding effect, whereas the employment of fixed kernels of a large scale introduces redundancy in computation. Given the significant variations in the scales of targets within authentic underwater scenes, a mechanism capable of the dynamic selection and aggregation of multi-scale receptive fields during the stage of shallow feature extraction is essential. Furthermore, the transmission of deep features lacks explicit channels for the preservation of boundaries. Existing networks for the fusion of multi-scale features, such as FPN and PANet, tend to smooth out the weak boundary signals of mimetic organisms or small targets after repeated operations of pooling and downsampling [14,42]. Conventional mechanisms of attention typically perform undifferentiated global or local weighting on feature maps, which fails to provide specialized refinement and the enhancement of sensitivity for the spatial boundary features of high frequency. Therefore, the achievement of adaptive perception for contextual receptive fields in the shallow layers of the network, and the assurance of precise refinement for blurred target boundaries before deep features enter the detection head, while models maintain the lightweight nature of one-stage detection, are critical for the enhancement of visual perception accuracy for AUVs in underwater environments.

These studies indicate that receptive field expansion and attention-based feature recalibration are useful for underwater perception. However, many existing designs mainly enhance feature responses at a general semantic level, while the distinction between shallow contextual insufficiency and boundary weakening before prediction is less explicitly modeled. In this work, these two issues are addressed separately: CALSK is employed to enhance context-aware shallow feature extraction, while SBAM is introduced to refine boundary-related responses prior to the detection head.

2.3. Bounding Box Regression and Loss Function Optimization

In one-stage object detection frameworks, the bounding box regression loss directly affects spatial localization accuracy. Early methods for detection were commonly employed the L_1 or L_2 norms to calculate the coordinate differences between predicted bounding boxes and ground truth boxes. However, this approach decouples the geometric correlation between individual coordinates and exhibits high sensitivity to the scales of targets. Consequently, the Intersection over Union (IoU) loss was introduced to improve the performance of regression through the inherent scale invariance of this metric. To address the issue of gradient vanishing when bounding boxes do not overlap, researchers have developed various optimizations. Rezatofighi et al. [43] proposed Generalized IoU (GIoU), which utilizes a minimum enclosing area to facilitate the optimization for non-overlapping boxes. Subsequently, Zheng et al. [44] introduced Distance IoU (DIoU) and Complete IoU (CIoU). CIoU incorporates three geometric factors, specifically the area of overlap, the normalized distance between centers, and the consistency of aspect ratios, into the penalty term. This integration accelerates the convergence of the model and establishes CIoU as the default loss function for mainstream one-stage detectors, which include YOLOv8. Nevertheless, the application of standard loss functions, such as CIoU, to complex underwater scenarios frequently results in an imbalance of optimization. As discussed in Section 2.1, real-world deep-sea datasets, such as RUOD, are characterized by a long-tailed distribution, resulting in substantial variations in visual identification difficulty across categories. For instance, underwater man-made facilities with regular geometric edges represent simple samples, whereas benthic organisms with mimetic features constitute difficult samples. The conventional CIoU loss assigns equal weights to all samples and lacks the capacity to perceive the difficulty of samples. During the backpropagation of gradients, the accumulated gradients from numerous simple samples and background negatives easily dominate the direction of updates for the network, which restricts the capabilities of localization for the model regarding targets of high difficulty, such as camouflaged organisms [45].

To mitigate the gradient allocation problem caused by sample imbalance, Tong et al. [46] introduced Wise-IoU (WIoU v3) as an improved bounding box regression loss for YOLOv8. This method reduces the dependence on aspect ratio and incorporates a dynamic non-monotonic focusing mechanism governed by the degree of outliers. WIoU functions by assessing the quality of predicted bounding boxes and subsequently diminishing the gradient contributions for anchor boxes of both exceptionally high and low quality. By concentrating optimization on samples of medium quality, the model prevents low-quality gradients from degrading overall generalization performance. Although WIoU demonstrates efficacy in general object detection tasks, its application in extreme underwater environments remains constrained. A primary limitation is that the method allocates weights based exclusively on the spatial quality of bounding boxes, which fails to account for the inherent difficulty of categories. Within complex underwater environments, predicted boxes with suboptimal geometric quality frequently represent severely occluded or rare difficult targets, rather than outliers resulting from erroneous annotations or stochastic noise. Consequently, the non-monotonic focusing mechanism in WIoU suppresses the backpropagation of gradients for these critical samples, which leads to a reduction in recall for difficult categories. Existing loss functions for bounding box regression fail to establish a coupled relationship between the degree of geometric outliers and the scarcity of categories. To ensure that baseline models, such as YOLOv13, can accommodate inter-class difficulty variations and long-tailed distribution characteristics of underwater organisms, a novel loss function is necessary. This function must evaluate the geometric quality of anchor boxes while capturing the intrinsic difficulty of detection for target categories to establish a

dual-path weighting mechanism. These considerations provide the theoretical foundation for the MCAWIoU loss proposed in this study.

Existing weighted IoU losses mainly adjust regression weights according to the geometric quality of individual predicted boxes. However, they usually do not explicitly consider whether a category maintains low localization quality over recent training iterations. MCAWIoU addresses this issue by using category-level EMA-IoU as an estimate of localization difficulty. Categories with lower recent localization quality receive relatively larger regression weights, which can alleviate localization imbalance associated with long-tailed underwater data.

3. Methodology

To make the overall methodology clearer, Figure 1 summarizes the processing flow of YOLO-CAB from underwater image input to detection output and training optimization.

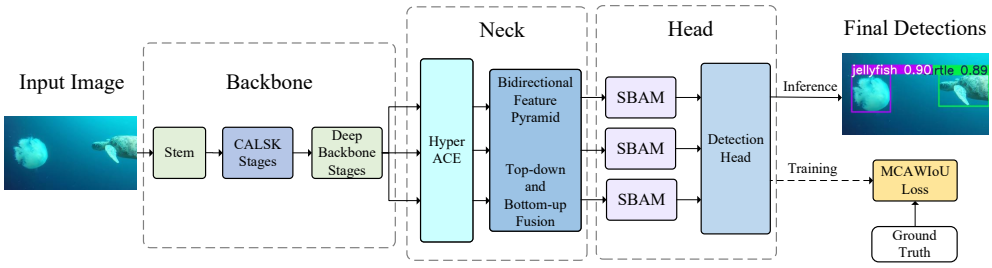

Figure 1. Overall methodological workflow of YOLO-CAB, showing the flow from underwater image input to backbone feature extraction, neck feature fusion, SBAM-enhanced detection, inference output, and MCAWIoU-based training optimization.

3.1. Overall Network Architecture of YOLO-CAB

In this study, YOLOv13 [47] is selected as the baseline model. As a state-of-the-art framework in one-stage object detection, it achieves an effective trade-off between detection precision and inference latency through a streamlined network topology and efficient parameter utilization. Architecturally, the vanilla YOLOv13 consists of three primary components: a feature extraction backbone, a feature fusion neck, and decoupled detection heads. Within the backbone, YOLOv13 extensively utilizes depthwise separable cross-stage partial (DSCSP) modules, which effectively reduce computational complexity through local feature aggregation. In the neck network, cross-layer connections are incorporated to facilitate the fusion of multi-scale features. This architectural design endows YOLOv13 with a high processing frame rate, making it more favorable for edge-oriented inference.

However, the direct deployment of the vanilla YOLOv13 architecture in complex underwater vision tasks reveals several inherent limitations. First, the standard convolutional modules in the shallow layers of the backbone utilize fixed kernel sizes and lack a mechanism for dynamic receptive field expansion. This makes it difficult to capture global contextual information for targets with significant scale variations in low-contrast scenarios. Furthermore, the coarse-grained feature fusion in the neck lacks an explicit mechanism to preserve high-frequency spatial information; consequently, the weak boundary features of camouflaged underwater targets are prone to attenuation during deep feature propagation. Finally, the native loss function in the detection head performs uniform optimization across all categories, failing to effectively address the pervasive long-tail distributions and sample difficulty imbalance characteristic of real-world seabed environments.

YOLO-CAB is designed around three related error sources in underwater detection: insufficient context under low contrast, weakened boundary cues after multi-scale feature propagation, and imbalanced localization optimization across categories. CALSK,

SBAM, and MCAWIoU are therefore placed in the backbone, the pre-head refinement stage, and the regression objective, respectively. This arrangement connects context-aware feature extraction, boundary-sensitive refinement, and category-aware localization optimization within the same detection pipeline.

To address the aforementioned issues, this work develops YOLO-CAB as a YOLOv13-based underwater object detector, as illustrated in Figure 2. The term YOLOv13n is used for the unmodified baseline in all comparisons and ablation experiments, while YOLO-CAB refers to the complete proposed detector after integrating CALSK, SBAM, MCAWIoU, and the restructured detection head. These modifications are organized according to the stages of the detection pipeline rather than introduced as isolated components. The resulting model improves underwater detection from three aspects: feature extraction, boundary perception, and localization optimization. Specifically, CALSK is integrated into the shallow layers of the backbone to strengthen the capabilities for the multi-scale capture of features. Furthermore, SBAM is designed and positioned prior to the input of the feature pyramid into the detection head, which refines multi-scale features and reinforces sensitivity to the boundaries of targets. Concurrently, the decoupled detection head is strategically restructured: four FullPAD Tunnel modules are removed, and the depthwise separable modules for the aggregation of features (DSC3k2) within the path of detection are replaced with standard C3k2 modules to reduce computational redundancy and improve the efficiency of feature decoding. Finally, the MCAWIoU loss is introduced at the output stage. By dynamically tracking the IoU at the category level, this loss function improves the precision of localization and the distribution of confidence for challenging underwater samples from the perspective of optimization.

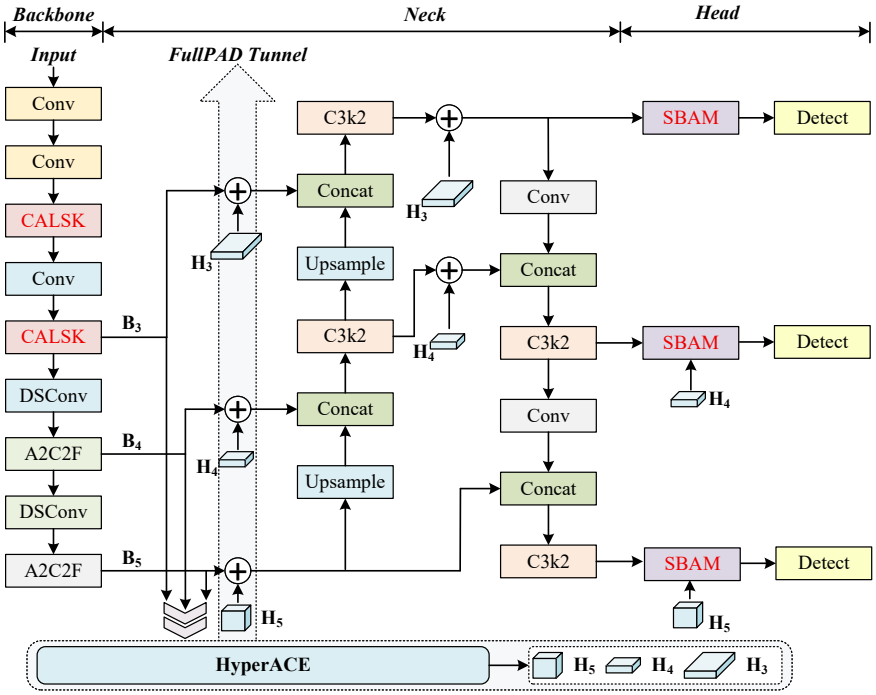

Figure 2. The overall network architecture of the proposed YOLO-CAB. The framework consists of a backbone integrated with the CALSK module, a neck for the fusion of features, and restructured detection heads equipped with the SBAM.

3.2. CALSK Module

Severe medium scattering and optical attenuation in real underwater environments cause low target–background contrast and significant scale variations. Under such low signal-to-noise ratio (SNR) conditions, the fixed local receptive fields in the shallow layers

of standard networks fail to capture global contextual correlations. To address this, we introduce CALSK into the backbone. By replacing standard convolutions in these shallow layers, CALSK expands the receptive field and adaptively weights critical spatial locations with controlled computational overhead, as illustrated in Figure 3.

This design is adopted because shallow underwater features require a larger contextual range while still preserving local texture details. The depthwise multi-scale structure enlarges the receptive field with limited computational increase, and the additive fusion avoids additional channel compression on low-contrast shallow features.

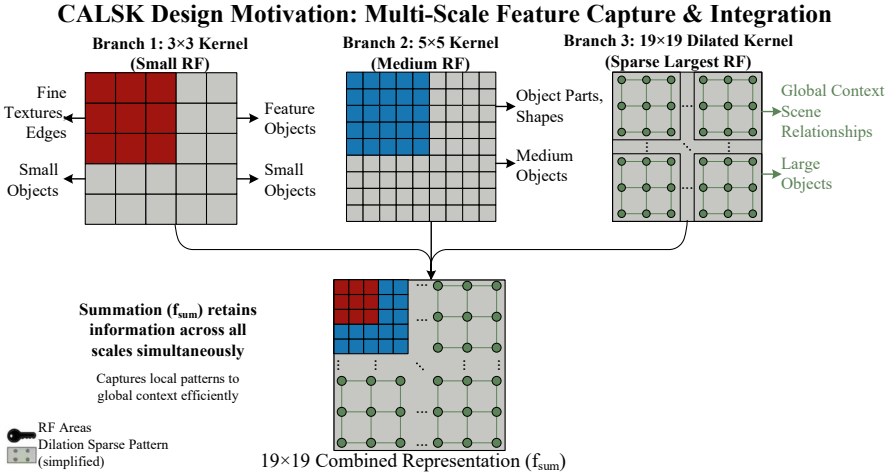

Figure 3. The structure of the CALSK module.

The multi-scale feature extraction process is formulated as:

$$\begin{aligned} U_1 &= \mathrm{DWConv}_{3\times3}(X), \\ U_2 &= \mathrm{DWConv}_{5\times5}(U_1), \\ U_3 &= \mathrm{DWConv}_{7\times7,d=3}(U_2) \end{aligned} \tag{1}$$

where X is the input feature map, $\mathrm{DWConv}_{k\times k}$ denotes a depthwise convolution with a $k \times k$ kernel, and $d = 3$ indicates a dilation rate of 3. This multi-scale depthwise convolution design forms the core of CALSK. By utilizing varying kernel dimensions and incorporating dilated convolutions for larger scales, this configuration expands the theoretical receptive field to capture features across multiple spatial scales. Instead of conventional channel concatenation followed by dimensionality reduction, CALSK employs element-wise addition across the three scales to obtain the fused feature U_{fused}. This strategy preserves the semantic information of each channel while preventing the loss of spatial structural details typically caused by dimensionality reduction.

The fused feature U_{fused} is subsequently fed into a spatial attention sub-module. We apply average pooling and max pooling along the channel dimension to extract dual-channel spatial descriptors, capturing global smoothing priors and high-frequency saliency features, respectively. These descriptors are concatenated and compressed via a 7×7 spatial convolution, followed by a Sigmoid activation function to generate a single-channel spatial mask M_s:

$$M_s = \sigma(\mathrm{Conv}_{7\times7}([\mathrm{AvgPool}(U_{\text{fused}}); \mathrm{MaxPool}(U_{\text{fused}})])) \tag{2}$$

where σ denotes the Sigmoid function, $\mathrm{Conv}_{7\times7}$ is the spatial convolution, and $[\cdot;\cdot]$ represents the concatenation operation. This mask acts as a contrast modulator and is applied directly to the fused features, adaptively weighting low-contrast regions of underwater

targets. Consequently, this operation enhances the feature representation of weak targets during the initial extraction stages.

3.3. SBAM

In one-stage object detection frameworks, multi-scale features are susceptible to the attenuation of subtle boundary signals from underwater benthic organisms, primarily caused by repeated deep pooling and downsampling operations. To refine these multi-scale features and enhance sensitivity to target boundaries, we introduce SBAM at the three output scales (P_3, P_4, and P_5) of the feature pyramid prior to the detection head, as illustrated in Figure 4.

SBAM is placed before the detection head because boundary cues may be weakened after repeated feature fusion and downsampling. The two-branch design is used to capture local edge responses and broader contour patterns separately before generating the spatial attention mask.

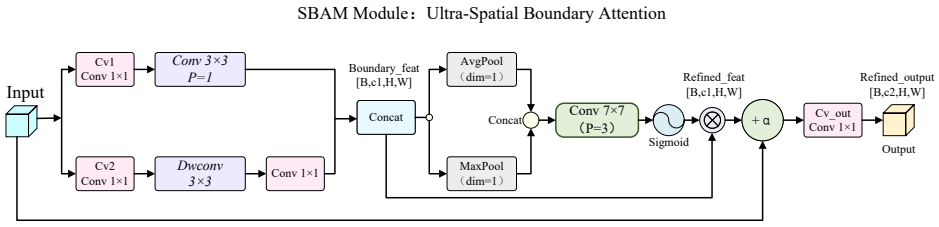

Figure 4. The detailed architecture of the SBAM.

SBAM explicitly isolates edge and contour information from fused multi-scale features. Given an input feature map F_{in}, the module first performs dimensionality reduction via two parallel 1×1 convolutional branches. These paths then enter different receptive field branches to extract high-frequency signals: the first branch uses a standard 3×3 convolution to capture local, compact edge cues, while the second employs a sequence of depthwise separable convolutions to capture broader geometric contours:

$$\begin{aligned} F_{\text{b1}} &= \text{Conv}_{3\times3}(\text{Conv}_{1\times1}(F_{\text{in}})), \\ F_{\text{b2}} &= \text{DWConv}_{\text{seq}}(\text{Conv}_{1\times1}(F_{\text{in}})) \end{aligned} \tag{3}$$

where $\text{Conv}_{k\times k}$ denotes a standard convolution and $\text{DWConv}_{\text{seq}}$ denotes the sequence of depthwise separable convolutions. The extracted features from both paths are concatenated into a unified boundary tensor $F_{\text{bound}} = [F_{\text{b1}}; F_{\text{b2}}]$. To prevent incorrect attention bias from under-fused features, SBAM calculates a spatial attention mask directly from this boundary tensor. We apply average and max pooling along the channel dimension, followed by spatial convolution and activation, to generate a contour-sensitive mask. This mask is then used for the element-wise weighted refinement of the boundary features:

$$M_{\text{s}} = \sigma(\text{Conv}_{7\times7}([\text{AvgPool}(F_{\text{bound}}); \text{MaxPool}(F_{\text{bound}})])) \tag{4}$$

where σ denotes the Sigmoid function and $[\cdot;\cdot]$ represents the concatenation operation. To stabilize gradient backpropagation and enable the network to adaptively adjust the boundary information injection based on varying underwater scenarios, we introduce a learnable dynamic factor α for a residual connection between the refined boundary features and the original input. Initialized to a small constant for early training stability, α is adaptively tuned via backpropagation, providing high-quality, boundary-enhanced features to the detection head:

$$F_{\text{out}} = \text{Conv}_{1\times1}(F_{\text{in}} + \alpha \cdot (M_{\text{s}} \otimes F_{\text{bound}})) \tag{5}$$

where $\otimes$ denotes element-wise multiplication.

3.4. MCAWIoU Loss

Conventional bounding box regression losses in YOLOv13 do not differentiate optimization weights across categories. Consequently, backpropagation is easily dominated by abundant common categories, reducing localization precision for rare and camouflaged organisms. Although the existing WIoU introduces a non-monotonic focusing mechanism, it relies exclusively on the geometric spatial quality of predicted boxes and lacks perception of inherent category difficulty. To address this, we propose the MCAWIoU loss.

The EMA-based category-level IoU is introduced to alleviate the localization imbalance caused by long-tailed category distributions. In underwater datasets, categories with fewer samples or ambiguous appearances may maintain lower localization quality during training, whereas frequent categories can dominate the regression gradients. By tracking the recent IoU of each category, MCAWIoU assigns relatively larger regression weights to categories with lower estimated localization quality. This design introduces category-level difficulty into the IoU regression objective without relying on manually fixed class weights.

Based on the non-monotonic focusing mechanism, MCAWIoU incorporates an Exponential Moving Average (EMA) to track IoU at the category level:

$$\mathrm{IoU}_c^{\mathrm{EMA}}(t) = \gamma \mathrm{IoU}_c^{\mathrm{EMA}}(t-1) + (1-\gamma)\mathrm{IoU}_c(t) \tag{6}$$

where γ serves as the momentum parameter, and $\mathrm{IoU}_c(t)$ denotes the average IoU of category c at iteration t. During training, the model continuously tracks the average IoU for each category to automatically identify difficult samples and achieve data-driven dynamic adjustments. The difficulty level for a category is defined as $D_c = 1 - \mathrm{IoU}_c^{\mathrm{EMA}}$. A higher D_c indicates greater prediction difficulty for that category at the current stage. Based on this real-time assessment, MCAWIoU constructs a dual-path weighting mechanism. In the IoU regression path, an adaptive base weight driven by category difficulty is introduced:

$$W_{\mathrm{box}}(c) = 1 + \beta \cdot \sigma\left(\frac{D_c}{\tau}\right) \tag{7}$$

where hyperparameters β and τ dictate the scale and sensitivity of this weight. To avoid training oscillations caused by hard boundary switching during anchor box weight calculation, MCAWIoU uses soft label scores to perform a weighted summation of the enhancement coefficients across all categories, generating a smoothed final weight:

$$W_{\mathrm{final}} = \sum_{c=1}^{C} \mathrm{Score}_c \cdot W_{\mathrm{box}}(c) \tag{8}$$

where Score_c indicates the predicted probability for category c, and C represents the total number of categories. This smoothed weight W_{final} is applied to the base loss, enabling the network to stably focus localization regression on difficult samples. The final regression loss, coupled with the non-monotonic focusing factor R_{wiou}, is defined as:

$$L_{\mathrm{MCAWIoU}} = W_{\mathrm{final}} \cdot R_{\mathrm{wiou}} \cdot L_{\mathrm{iou}} \tag{9}$$

where L_{iou} corresponds to the standard IoU loss.

In the classification path, a similar mechanism weights the Binary Cross-Entropy (BCE) loss to optimize the confidence distribution. To prevent excessive classification optimization from increasing false positives, we impose differentiated constraints on the enhancement magnitudes for both paths. This synergistic dual-path weighting and soft label smoothing

ensures loss gradient continuity, effectively mitigating missed detections and false positives for long-tail underwater samples.

4. Experiments

4.1. Dataset Description

In this section, we evaluate the proposed approach on the Real-world Underwater Object Detection (RUOD) dataset. Unlike early datasets collected in controlled pools or single nearshore environments, the RUOD dataset contains images sourced from diverse maritime regions under varying lighting conditions. It captures physical degradations typical of deep-sea environments, such as color shifts, low visibility, and light scattering (illustrated in Figure 5).

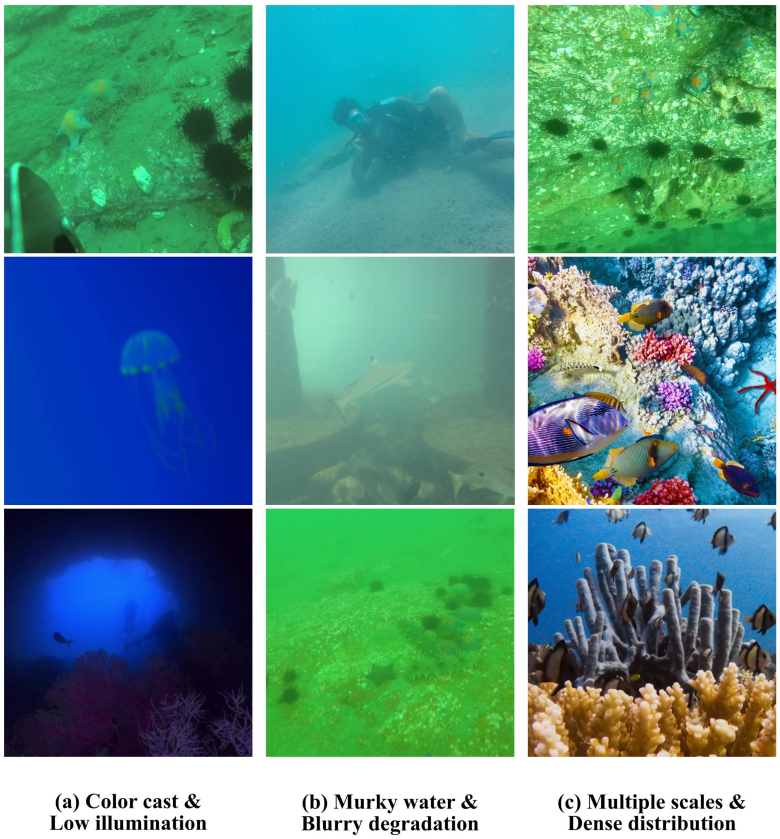

Figure 5. Challenges of visual degradation in authentic underwater environments: (**a**) color cast and low illumination, (**b**) murky water and blurry degradation, and (**c**) multiple scales and dense distribution.

The dataset includes ten target categories: holothurian (sea cucumber), echinus (sea urchin), scallop, starfish, fish, corals, divers, cuttlefish, turtles, and jellyfish, as shown in Figure 6. These categories exhibit a pronounced long-tail distribution, visualized in Figure 7. Additionally, many benthic organisms (e.g., sea cucumbers and scallops) rely on strong environmental camouflage, with their boundary contours blending seamlessly into the seabed. This data imbalance and low signal-to-noise ratio (SNR) rigorously test the model's capacity to localize difficult samples and optimize class balance. We partition the dataset into training, validation, and testing sets using a 7:2:1 ratio, yielding 9800 training images, 2800 validation images, and 1400 testing images.

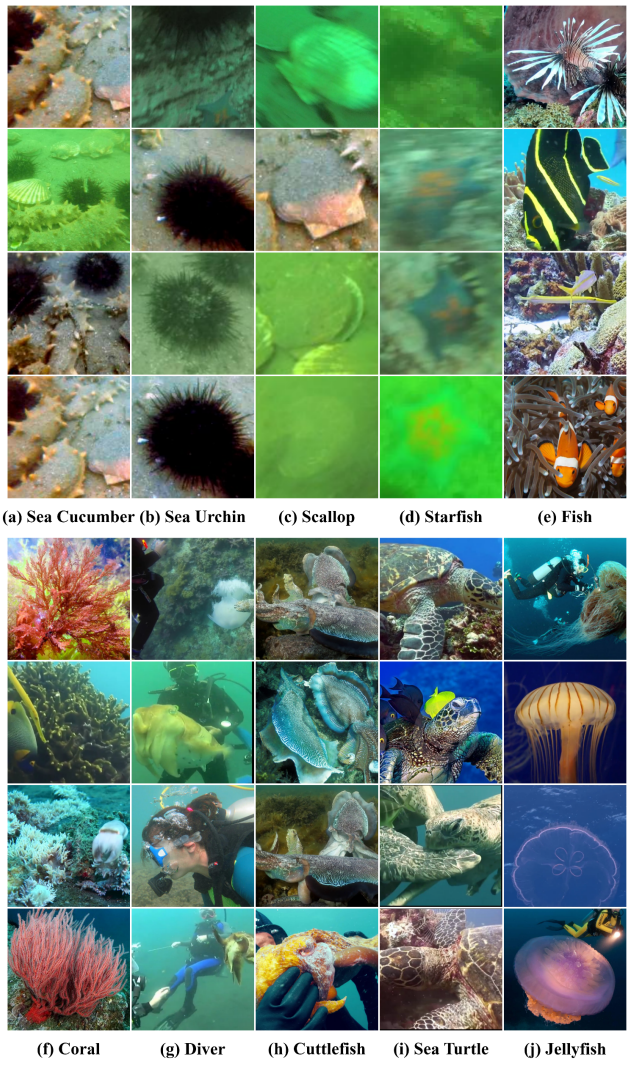

Figure 6. Typical examples of the 10 target categories in the RUOD dataset.

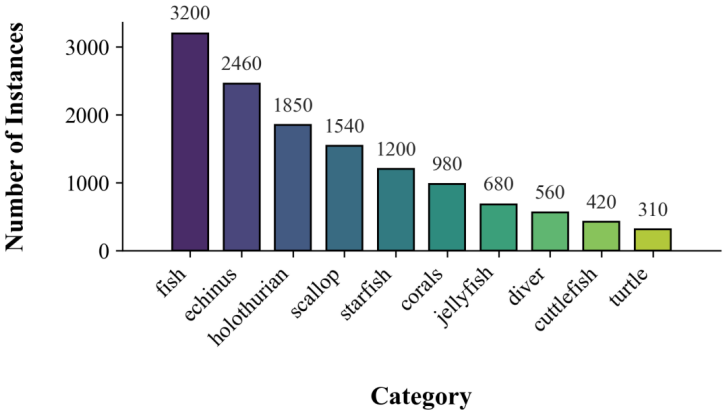

Figure 7. Thedata distribution of the 10 categories in the RUOD dataset, exhibiting a pronounced long-tail characteristic.

4.2. Experimental Setup

4.2.1. Training Details

Experiments were implemented in PyTorch version 2.1.2 and conducted on a server equipped with an NVIDIA GeForce RTX 4090 GPU (24 GB VRAM). During training, input images were resized to a resolution of 640 × 640 pixels. We optimized the model using Stochastic Gradient Descent (SGD) with an initial learning rate of 0.01, modulated by a

cosine annealing scheduler. The batch size was set to 64. Training lasted for 300 epochs, including a 5-epoch warmup period. Data augmentation strategies included Mosaic (1.0) and Mixup (0.1). To preserve the spatial distribution of authentic data during the late training stages, Mosaic augmentation was disabled for the final 10 epochs.

4.2.2. Evaluation Metrics

We evaluate detection performance using Precision (P), Recall (R), F1-score ($F1$), Mean Average Precision at an IoU threshold of 0.5 (mAP_{50}), and $mAP_{50–95}$. To measure computational efficiency, we report parameter count (K), floating-point operations (GFLOPs), and inference time (ms).

Precision measures the accuracy of positive predictions:

$$P = \frac{TP}{TP + FP} \tag{10}$$

where TP and FP denote the number of true positives and false positives, respectively. Recall reflects the proportion of actual positive samples correctly identified:

$$R = \frac{TP}{TP + FN} \tag{11}$$

where FN represents the number of false negatives. The F1-score is the harmonic mean of Precision and Recall, providing a balanced assessment for imbalanced datasets:

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{12}$$

The mean Average Precision (mAP) averages the Average Precision (AP) across all n categories:

$$mAP = \frac{1}{n} \sum_{j=1}^{n} AP_j \tag{13}$$

specifically, mAP_{50} evaluates performance at an IoU threshold of 0.5, while $mAP_{50–95}$ averages the mAP across IoU thresholds from 0.5 to 0.95 in increments of 0.05, thereby imposing stricter localization requirements.

4.3. Comparative Experiments

4.3.1. Overall Performance Comparison

Table 1 lists the overall performance comparison between YOLO-CAB and several object detectors on the RUOD dataset. In addition to YOLOv8n, YOLOv10n, YOLOv11n, YOLOv12n, and the vanilla YOLOv13n baseline, we include EAW-YOLO11 [48], a representative underwater-oriented detector, to provide a more relevant comparison for degraded underwater imagery, ambiguous boundaries, and category imbalance.

Table 1. The overall comparison of performance on the RUOD dataset.

Methods	mAP_{50}	$mAP_{50–95}$	Precision	Recall	F1-Score
YOLOv8n	77.2	52.1	81.2	69.5	74.9
YOLOv10n	77.8	52.9	81.4	70.4	75.5
YOLOv11n	77.6	52.6	82.2	69.9	75.5
YOLOv12n	77.1	52.7	81.4	69.7	75.1
EAW-YOLO11	78.2	53.8	82.0	70.8	76.0
Baseline	76.3	52.1	81.2	68.5	74.3
Ours	**81.0**	**55.1**	**83.0**	**3.6**	**78.0**

Note: Bold values indicate the best performance for each metric.

As shown in Table 1, YOLO-CAB achieves the best values among the compared methods for all reported accuracy metrics. Compared with the YOLOv13n baseline, it improves mAP_{50} by 4.67%, $mAP_{50–95}$ by 3.0%, and F1-score by 3.7%. Compared with EAW-YOLO11, YOLO-CAB improves mAP_{50} by 2.8%, $mAP_{50–95}$ by 1.3%, and F1-score by 2.0%, which provides a more targeted reference for challenging underwater scenes. These gains are meaningful because underwater images often contain low contrast, background interference, and blurred target boundaries. They are also consistent with prior studies showing that attention-based multi-scale context modeling and selective kernel learning improve detection robustness by adapting receptive fields and suppressing background interference in complex visual scenes [33,34,49]. In YOLO-CAB, CALSK enlarges the effective receptive field in shallow layers, which helps retain discriminative cues for low-contrast underwater targets, while SBAM refines boundary-sensitive responses before prediction. This design is further supported by studies on feature pyramids and boundary-preserving representation learning, which indicate that repeated pyramid fusion may weaken fine spatial details and that explicit boundary modeling can improve localization quality [14,42,50]. Thus, the performance gains are consistent with the intended roles of context aggregation and boundary-aware feature refinement. The convergence curves for Precision, Recall, and mAP are compared with the baseline in Figure 8. A revised qualitative comparison between YOLO-CAB and other YOLO variants is presented in Figure 9, where representative underwater scenarios are annotated to make the cross-model differences easier to inspect.

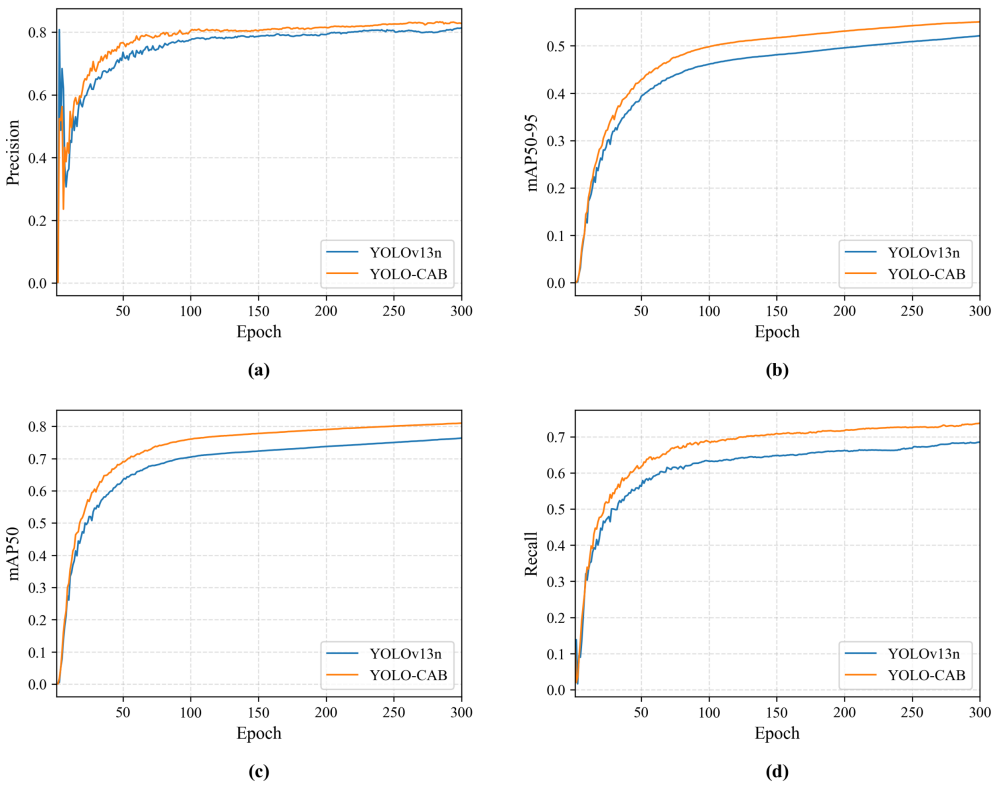

Figure 8. Training performance comparison between the vanilla YOLOv13n baseline and YOLO-CAB: (**a**) Precision, (**b**) mAP@0.5:0.95, (**c**) mAP@0.5, and (**d**) Recall.

Despite the overall improvement brought by YOLO-CAB, several failure modes remain under highly challenging underwater conditions. When visibility becomes extremely poor, severe attenuation weakens target appearance and suppresses boundary cues, which makes reliable localization difficult. When foreground objects exhibit strong visual similarity to the surrounding background, the model may still confuse target regions with

background structures. Dense overlap or partial occlusion can further hinder the separation of adjacent instances. In addition, very small objects remain difficult because only limited discriminative cues are preserved at low spatial resolution. These observations indicate that CALSK, SBAM, and MCAWIoU improve robustness, but they do not fully remove the difficulty caused by extreme degradation, cluttered scenes, and tiny-object perception. Representative failure cases are shown in Figure 10.

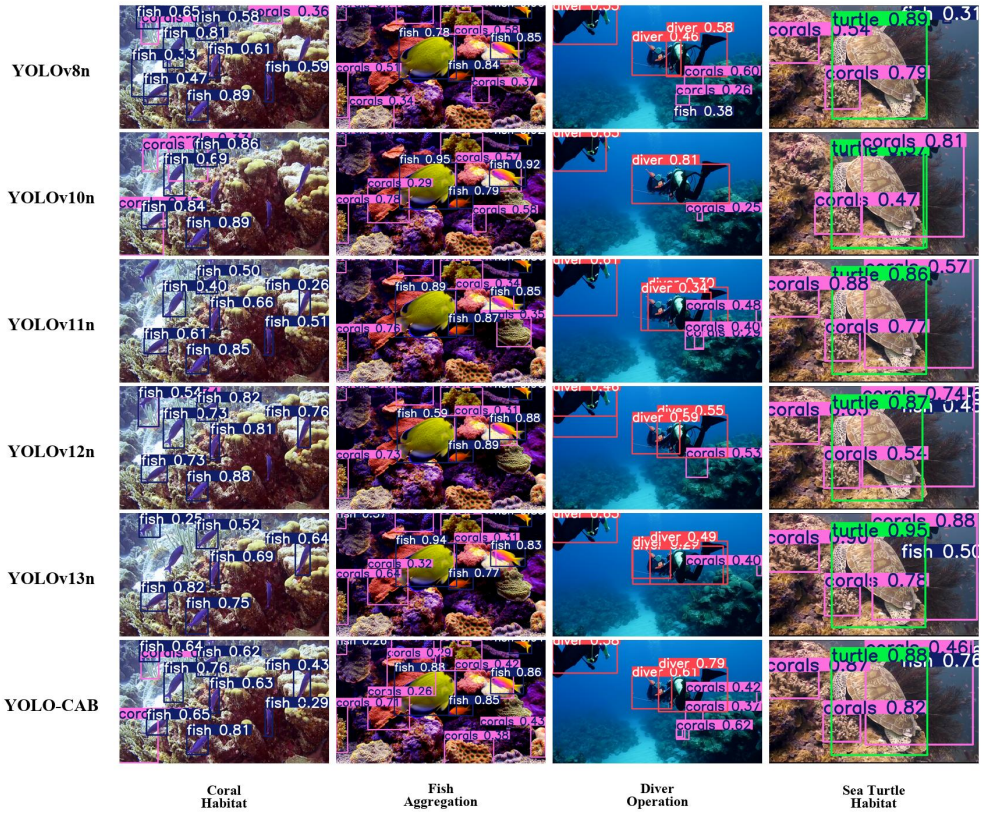

Figure 9. A qualitative comparison of detection results between several YOLO variants and YOLO-CAB in representative challenging underwater scenes, including coral habitat, fish aggregation, diver operation, and sea turtle habitat scenarios.

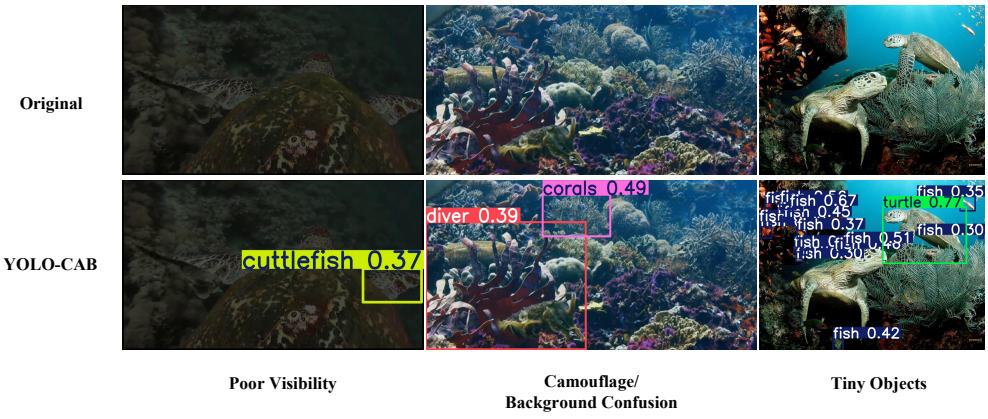

Figure 10. Representative failure cases of YOLO-CAB under challenging underwater conditions. From left to right, the examples correspond to poor visibility, camouflage-induced background confusion, and tiny-object scenes.

4.3.2. Detection Performance Comparison Across Categories

To further evaluate category-wise detection performance, Table 2 compares the mAP_{50} values for different methods across the 10 RUOD categories.

Table 2 reports the category-wise mAP_{50}, while Figure 7 shows the long-tail distribution of the RUOD dataset. YOLO-CAB achieves relatively high accuracy for Sea Turtle, Cuttlefish, Diver, Sea Urchin, and Starfish, which usually have clearer global shape cues or stronger object-background separation. In contrast, Sea Cucumber, Scallop, Fish, and Coral obtain larger relative gains but remain more difficult overall because they are more strongly affected by long-tail imbalance, camouflage, small effective target regions, or blurred and irregular boundaries. Jellyfish remains the weakest case among the ten categories, as its translucent appearance and diffuse contour reduce boundary reliability and make classification more sensitive to noise and background interference. These observations suggest that CALSK and SBAM improve robustness to degraded boundaries, while MCAWIoU helps reduce optimization bias toward difficult and underrepresented categories. Nevertheless, boundary ambiguity and sample imbalance still constrain the hardest categories.

Table 2. A comparison of mAP_{50} for each category on the RUOD dataset.

Category	YOLOv8n	YOLOv10n	YOLOv11n	YOLOv12n	Baseline	Ours
Sea Cucumber	65.8	65.2	65.4	64.3	61.7	**71.1**
Sea Urchin	87.7	89.2	88.2	88.1	86.9	**90.9**
Scallop	64.9	65.8	63.0	62.6	63.6	**73.0**
Starfish	83.9	83.8	82.8	82.6	82.7	**86.2**
Fish	67.6	67.8	68.6	68.5	67.4	**73.1**
Coral	62.8	64.9	65.1	63.1	63.9	**68.5**
Diver	88.4	88.8	88.6	89.2	89.4	**91.3**
Cuttlefish	92.8	92.6	92.6	93.2	91.7	**94.6**
Sea Turtle	94.1	93.3	94.5	94.6	94.2	**95.2**
Jellyfish	64.5	66.2	**66.9**	65.2	62.0	65.8

Note: Bold values indicate the best performance for each category.

4.3.3. Cross-Dataset Validation on DUO

To further evaluate generalization beyond RUOD, we conduct an additional validation on the DUO dataset [25] under a shared-category protocol. As shown in Table 3, YOLO-CAB maintains the best overall performance, reaching 77.6 in mAP_{50}, 55.2 in $mAP_{50–95}$, 80.9 in Precision, 70.7 in Recall, and 75.4 in F1-score. Compared with the strongest baseline, YOLOv8n, YOLO-CAB improves mAP_{50} by 2.2%, $mAP_{50–95}$ by 0.8%, and F1-score by 2.5%. These results provide additional evidence that the proposed method remains effective across underwater datasets with different data distributions and visual characteristics under the same training setting.

Table 3. Cross-dataset validation on the DUO dataset under the shared-category protocol.

Methods	mAP_{50}	$mAP_{50–95}$	Precision	Recall	F1-Score
YOLOv8n	75.4	54.4	79.7	67.1	72.9
YOLOv10n	74.1	52.9	78.2	66.4	71.8
YOLOv11n	75.2	54.1	79.1	68.2	73.2
YOLOv12n	73.4	52.3	78.0	67.5	72.4
YOLOv13n Baseline	73.4	51.7	78.4	66.2	71.8
Ours (YOLO-CAB)	**77.6**	**55.2**	**80.9**	**70.7**	**75.4**

Note: Bold formatting indicates the best-performing method and the best value for each metric.

4.4. Ablation Study

To examine the contribution of each proposed component to YOLO-CAB, this section uses vanilla YOLOv13n as the baseline and progressively adds CALSK, SBAM, and MCAWIoU. The results are summarized in Table 4.

The ablation results further clarify why the performance improves. Adding CALSK alone increases mAP_{50} from 76.3% to 78.2%, which supports the role of adaptive large-kernel context modeling in handling scale variation and low contrast. This observation is consistent with multi-scale attention and selective kernel learning, where receptive-

field adaptation strengthens visual representation in complex scenes [33,34,49]. Adding SBAM alone improves mAP_{50} to 78.1% and $mAP_{50–95}$ to 53.9%, suggesting that explicit boundary refinement improves localization by compensating for the loss of weak contours during pyramid-based feature propagation [14,42,50]. Adding MCAWIoU alone raises mAP_{50} to 78.0%, but $mAP_{50–95}$ decreases from 52.1% to 51.8%. This result indicates that category-aware reweighting is most effective when supported by sufficient contextual and boundary features. The CALSK+SBAM variant achieves the strongest high-threshold localization, and the further incorporation of MCAWIoU raises mAP_{50} and Precision to their best values. These results show that the three components provide complementary benefits by jointly improving context modeling, boundary-sensitive localization, and category-aware optimization. This behavior is consistent with long-tailed detection studies showing that reweighted optimization changes gradient allocation and can favor difficult categories while interacting with localization quality [21,51,52].

Table 4. The study of ablation for different modules on the RUOD dataset.

Methods	mAP_{50}	$mAP_{50–95}$	Precision	Parameters (K)	GFLOPs
Baseline	76.3	52.1	81.2	**2449.8**	**6.13**
+SBAM	78.1	53.9	81.9	2873.9	6.78
+CALSK	78.2	53.9	82.5	2478.6	6.73
+MCAWIoU	78.0	51.8	81.7	**2449.8**	**6.13**
+SBAM+MCAWIoU	79.4	52.5	82.6	2873.9	6.78
+CALSK+MCAWIoU	79.3	53.2	82.8	2478.6	6.73
+CALSK+SBAM	79.8	**55.4**	82.1	2902.6	7.38
Ours (All)	**81.0**	55.1	**83.0**	3291.5	8.85

Note: Bold values indicate the best result in each column. For mAP and Precision, higher values are better; for Parameters and GFLOPs, lower values are better. The bold method name indicates the proposed complete model.

We next compare the CALSK+SBAM variant with the CALSK+SBAM+MCAWIoU variant at the category level, as shown in Figure 11. The gains are concentrated in several difficult categories rather than being evenly distributed across all classes. In particular, the scallop category improves by 5.32% in AP_{50}, 1.96% in $AP_{50–95}$, and 6.98% in Recall, while the corals category improves by 1.81% in AP_{50}, 0.81% in $AP_{50–95}$, and 3.29% in Recall. Averaged over the five most difficult categories under the CALSK+SBAM setting, adding MCAWIoU increases AP_{50} by 1.53% and Recall by 1.61%. In contrast, the three easiest categories show an average decrease of 1.31% in $AP_{50–95}$. This pattern indicates that MCAWIoU mainly reallocates optimization emphasis toward hard categories rather than providing uniform gains across all classes.

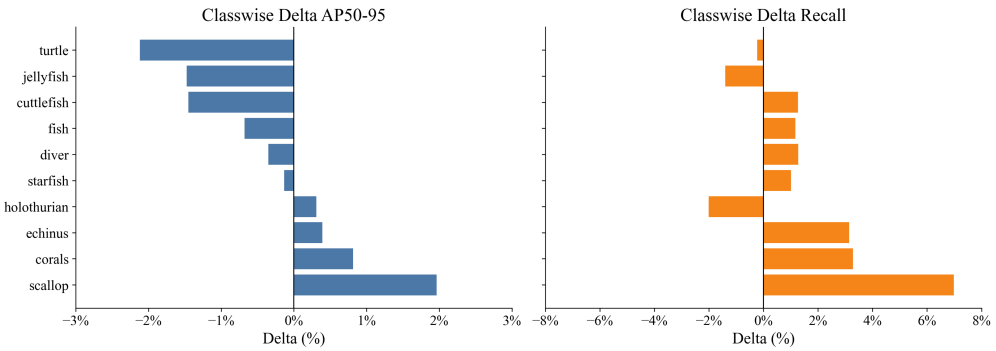

Figure 11. Category-specific gains introduced by MCAWIoU over the CALSK+SBAM variant.

We also track the per-class EMA IoU and the corresponding adaptive weights during training, as shown in Figure 12. The mean difficulty over all categories decreases from 0.6875 at epoch 1 to 0.1881 at epoch 300, and the mean IoU-branch weight decreases from 1.5500 to 1.1505, indicating that the reweighting effect gradually weakens as training

converges. Categories with persistently lower EMA IoU retain larger weights in the late stage. Over epochs 250–300, holothurian and scallop have mean difficulties of 0.2831 and 0.2826, corresponding to mean IoU-branch weights of 1.2265 and 1.2261 and mean classification weights of 1.1132 and 1.1130, whereas turtle has a mean difficulty of only 0.0648, with mean IoU-branch and classification weights of 1.0518 and 1.0259. These trends show that MCAWIoU adaptively emphasizes hard categories during training instead of relying on fixed manual class weights.

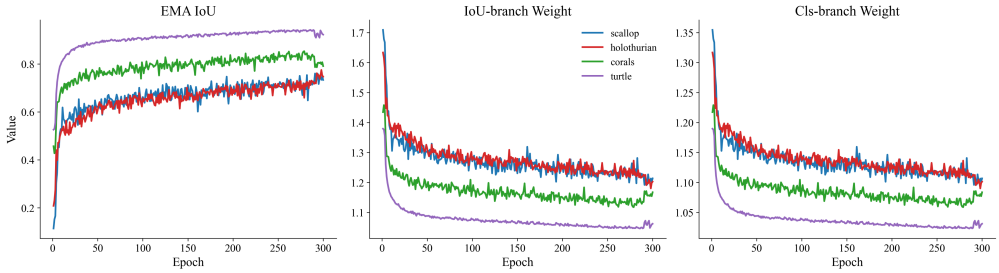

Figure 12. Training dynamics of category-aware EMA IoU and adaptive weights in MCAWIoU for representative categories.

4.5. Computational Efficiency Analysis

Table 5 presents the computational efficiency comparison among different methods, including EAW-YOLO11. The testing platform is a server equipped with an NVIDIA GeForce RTX 4090 GPU, and the input image resolution is set to 640 × 640. The reported inference time is the average value obtained from multiple evaluations.

Table 5. Computational efficiency comparison among different methods.

Methods	Parameters (K)	GFLOPs	Running Time (ms)	FPS
YOLOv8n	3007.6	8.09	**2.19**	**456.6**
YOLOv10n	3175.8	8.63	2.76	362.3
YOLOv11n	2584.1	6.32	2.50	400.0
YOLOv12n	2510.3	**5.83**	4.88	204.9
EAW-YOLO11	2607.5	6.40	4.70	212.9
Baseline	**2449.8**	6.13	4.41	226.8
Ours	3291.5	8.85	5.21	191.9

Note: Bold values indicate the best result in each column. For Parameters, GFLOPs, and Running Time, lower values are better; for FPS, higher values are better. The bold method name indicates the proposed method.

As shown in Table 5, EAW-YOLO11 contains 2607.5 K parameters and 6.40 GFLOPs, with a running time of 4.70 ms and a throughput of 212.9 FPS. YOLO-CAB introduces higher computational cost than the YOLOv13n baseline and EAW-YOLO11, but it delivers the strongest detection accuracy among all compared methods. These results indicate that the proposed method improves underwater detection performance while maintaining practical real-time efficiency under the tested hardware setting.

Table 6 provides a direct comparison of the trade-off between detection accuracy and computational efficiency for the YOLOv13n baseline and YOLO-CAB. Compared with the baseline, YOLO-CAB improves mAP_{50} by 4.7% and $mAP_{50–95}$ by 3.0%, while GFLOPs increase from 6.13 to 8.85 and inference time increases from 4.41 ms to 5.21 ms. Although the proposed model introduces additional computation, the absolute increase in latency is only 0.80 ms, and the model still runs at 191.9 FPS under the tested hardware setting. These results indicate that the gain in detection accuracy is achieved with a moderate efficiency cost and remains suitable for real-time inference in this experimental environment.

Table 6. Accuracy-efficiency trade-off between the YOLOv13n baseline and YOLO-CAB on the RUOD dataset. Inference time is measured on an NVIDIA GeForce RTX 4090 with an input size of 640 × 640.

Methods	mAP_{50}	$mAP_{50–95}$	GFLOPs	Running Time (ms)	FPS
Baseline	76.3	52.1	6.13	4.41	226.8
YOLO-CAB	81.0	55.1	8.85	5.21	191.9

To complement the desktop-GPU measurements, we further evaluate YOLO-CAB on an NVIDIA Jetson Orin Nano edge platform, as shown in Figure 13. The deployment-oriented inference results are summarized in Table 7. Under the tested settings, YOLO-CAB achieves 26.5 FPS in FP32 mode and 38.6 FPS in FP16 mode on the Orin Nano, where FPS is computed as 1000 divided by the inference time in milliseconds. The FP16 setting reduces latency from 37.7 ms to 25.9 ms while maintaining comparable detection accuracy, with 80.6% mAP_{50} and 54.2% $mAP_{50–95}$. These results provide additional evidence that YOLO-CAB can support real-time inference on Jetson-class hardware under the evaluated platform and precision settings.

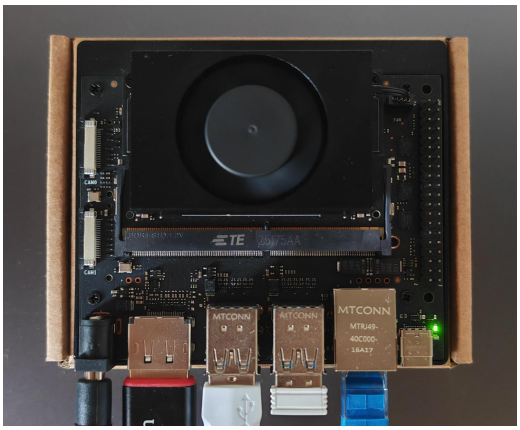

Figure 13. Embedded hardware platform used for deployment-oriented inference evaluation on NVIDIA Jetson Orin Nano.

Table 7. Embedded-platform inference performance of YOLO-CAB on NVIDIA Jetson Orin Nano, with the RTX 4090 result listed as a reference.

Platform	Mode	Precision	Recall	mAP_{50}	$mAP_{50–95}$	Inference (ms)	FPS
RTX 4090	FP32	83.0	73.6	81.0	55.1	5.21	191.9
Jetson Orin Nano	FP32	82.6	73.7	80.6	54.2	37.7	26.5
Jetson Orin Nano	FP16	82.5	73.7	80.6	54.2	25.9	38.6

5. Conclusions

In this paper, we propose YOLO-CAB, a YOLOv13-based underwater object detector designed for AUV visual perception under optical degradation, blurred boundaries, and long-tailed category distributions. By coordinating CALSK, SBAM, MCAWIoU, and detection-head restructuring, the method improves contextual feature extraction, boundary-sensitive localization, and category-aware optimization. Experiments on the RUOD dataset show that YOLO-CAB achieves 81.0% mAP_{50} and 55.1% $mAP_{50–95}$, improving the YOLOv13n baseline by 4.67% and 3.0%, respectively. Additional validation on the DUO dataset further supports its cross-dataset effectiveness under the shared-category protocol. The model also maintains a practical accuracy–efficiency trade-off, with 8.85 GFLOPs and an inference time of 5.21 ms under the tested hardware setting.

These results indicate that YOLO-CAB is effective for underwater object detection scenarios where low contrast, weak boundaries, and category imbalance jointly affect perception performance. The Jetson Orin Nano experiment also shows real-time inference on embedded hardware under the tested settings. Future work will extend the framework to multi-object tracking, integrate visual detection with sonar data, validate the method through sea trials, and improve cross-domain adaptation across underwater sensors, scenes, and category distributions.

Author Contributions: Conceptualization, R.L.; Methodology, R.L.; Software, Z.L.; Validation, S.B.; Formal analysis, R.L.; Resources, Z.L.; Writing—original draft, S.B.; Writing—review & editing, C.Y. and S.B.; Visualization, J.Y.; Supervision, C.Y. and J.Y.; Project administration, C.Y.; Funding acquisition, C.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 52401362, in part by the Natural Science Foundation of Liaoning Province under Grant 2025-BS-0231, and in part by the Dalian Science and Technology Talent Innovation Support Program under Grant 2025RQ31.

Data Availability Statement: The RUOD dataset utilized in this study is openly available in the GitHub repository at https://github.com/dlut-dimt/RUOD (accessed on 25 May 2026). Detailed characteristics and specifications of the dataset can be found in the associated publication at https://doi.org/10.1016/j.neucom.2022.10.039 (accessed on 25 May 2026).

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

References

1. Liu, F.; Xu, W.; Feng, Z.; Yu, C.; Liang, X.; Su, Q.; Gao, J. Task Allocation and Path Planning Method for Unmanned Underwater Vehicles. *Drones* **2025**, *9*, 411. [CrossRef]
2. Chen, Q.; Liu, B.; Yu, C.; Yang, M.; Guo, H. Task Allocation and Saturation Attack Approach for Unmanned Underwater Vehicles. *Drones* **2025**, *9*, 115. [CrossRef]
3. Sahoo, A.; Dwivedy, S.K.; Robi, P. Advancements in the field of autonomous underwater vehicle. *Ocean Eng.* **2019**, *181*, 145–160. [CrossRef]
4. Zereik, E.; Bibuli, M.; Mišković, N.; Ridao, P.; Pascoal, A. Challenges and future trends in marine robotics. *Annu. Rev. Control* **2018**, *46*, 350–368. [CrossRef]
5. Tang, S.; Qian, Y.; Guo, S.; Liu, Y.; Shen, H.; Yu, C. DSB-net: An effective dynamic sparse Bayesian network for underwater sonar object detection. *Neurocomputing* **2026**, 133288. [CrossRef]
6. Fan, B.; Chen, W.; Cong, Y.; Tian, J. Dual refinement underwater object detection network. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 275–291.
7. Bonin-Font, F.; Ortiz, A.; Oliver, G. Visual navigation for mobile robots: A survey. *J. Intell. Robot. Syst.* **2008**, *53*, 263–296. [CrossRef]
8. Jaffe, J.S. Computer modeling and the design of optimal underwater imaging systems. *IEEE J. Ocean. Eng.* **2002**, *15*, 101–111. [CrossRef]
9. Akkaynak, D.; Treibitz, T. Sea-thru: A method for removing water from underwater images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2019; pp. 1682–1691.
10. Ancuti, C.O.; Ancuti, C.; De Vleeschouwer, C.; Bekaert, P. Color balance and fusion for underwater image enhancement. *IEEE Trans. Image Process.* **2017**, *27*, 379–393. [CrossRef] [PubMed]
11. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef]
12. Huang, J.; Rathod, V.; Sun, C.; Zhu, M.; Korattikara, A.; Fathi, A.; Fischer, I.; Wojna, Z.; Song, Y.; Guadarrama, S.; et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2017; pp. 7310–7311.
13. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2016; pp. 779–788.
14. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2017; pp. 2117–2125.

15. Anwar, S.; Li, C. Diving deeper into underwater image enhancement: A survey. *Signal Process. Image Commun.* **2020**, *89*, 115978. [CrossRef]
16. Islam, M.J.; Xia, Y.; Sattar, J. Fast underwater image enhancement for improved visual perception. *IEEE Robot. Autom. Lett.* **2020**, *5*, 3227–3234. [CrossRef]
17. Wang, Y.; Guo, J.; Gao, H.; Yue, H. UIEC 2-Net: CNN-based underwater image enhancement using two color space. *Signal Process. Image Commun.* **2021**, *96*, 116250. [CrossRef]
18. Zhang, Y.; Liu, F.; Lyu, J.; Wei, Y.; Yu, C. HMPNet: A Feature Aggregation Architecture for Maritime Object Detection from a Shipborne Perspective. In *Proceedings of the 2025 IEEE International Conference on Multimedia and Expo (ICME)*; IEEE: New York, NY, USA, 2025; pp. 1–6.
19. Gao, J.; Zhang, Y.; Geng, X.; Tang, H.; Bhatti, U.A. PE-Transformer: Path enhanced transformer for improving underwater object detection. *Expert Syst. Appl.* **2024**, *246*, 123253. [CrossRef]
20. Guo, L.; Liu, X.; Ye, D.; He, X.; Xia, J.; Song, W. Underwater object detection algorithm integrating image enhancement and deformable convolution. *Ecol. Inform.* **2025**, *89*, 103185. [CrossRef]
21. Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; Yu, S.X. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2019; pp. 2537–2546.
22. Le, T.N.; Cao, Y.; Nguyen, T.C.; Le, M.Q.; Nguyen, K.D.; Do, T.T.; Tran, M.T.; Nguyen, T.V. Camouflaged instance segmentation in-the-wild: Dataset, method, and benchmark suite. *IEEE Trans. Image Process.* **2021**, *31*, 287–300. [CrossRef]
23. Wang, N.; Wang, Y.; Er, M.J. Review on deep learning techniques for marine object recognition: Architectures and algorithms. *Control Eng. Pract.* **2022**, *118*, 104458. [CrossRef]
24. Fisher, R.B.; Shao, K.T.; Chen-Burger, Y.H. Overview of the fish4knowledge project. In *Fish4Knowledge: Collecting and Analyzing Massive Coral Reef Fish Video Data*; Springer: Cham, Switzerland, 2016; pp. 1–17.
25. 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)*; IEEE: New York, NY, USA, 2021; pp. 1–6.
26. Pedersen, M.; Lehotskỳ, D.; Nikolov, I.; Moeslund, T.B. Brackishmot: The brackish multi-object tracking dataset. In *Proceedings of the Scandinavian Conference on Image Analysis*; Springer: Cham, Switzerland, 2023; pp. 17–33.
27. Hong, J.; Fulton, M.; Sattar, J. Trashcan: A semantically-segmented dataset towards visual detection of marine debris. *arXiv* **2020**, arXiv:2007.08097.
28. 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]
29. Moniruzzaman, M.; Islam, S.M.S.; Bennamoun, M.; Lavery, P. Deep learning on underwater marine object detection: A survey. In *Proceedings of the International Conference on Advanced Concepts for Intelligent Vision Systems*; Springer: Cham, Switzerland, 2017; pp. 150–160.
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2016; pp. 770–778.
31. Yu, C.; Gu, X.; Yao, Y.; Wang, S.; Liang, X.; Pan, L.; Xiao, Y. Real-time 6-DoF pose estimation for UUVs based on advanced CNN architecture with adaptive loss constraints. *Ocean Eng.* **2025**, *333*, 121425. [CrossRef]
32. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 21–37.
33. Yu, C.; Yin, H.; Rong, C.; Zhao, J.; Liang, X.; Li, R.; Mo, X. YOLO-MRS: An efficient deep learning-based maritime object detection method for unmanned surface vehicles. *Appl. Ocean Res.* **2024**, *153*, 104240. [CrossRef]
34. Qin, X.; Yu, C.; Liu, B.; Zhang, Z. YOLO8-FASG: A high-accuracy fish identification method for underwater robotic system. *IEEE Access* **2024**, *12*, 73354–73362. [CrossRef]
35. Fan, D.P.; Ji, G.P.; Sun, G.; Cheng, M.M.; Shen, J.; Shao, L. Camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2020; pp. 2777–2787.
36. Ding, X.; Zhang, X.; Han, J.; Ding, G. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2022; pp. 11963–11975.
37. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
38. Yin, H.; Yu, C.; Wu, C.; Dai, K.; Shi, J.; Xu, Y.; Zhu, Y. Swin Transformer-based maritime objects instance segmentation with dual attention and multi-scale fusion. *Comput. Vis. Image Underst.* **2025**, *262*, 104556. [CrossRef]
39. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2018; pp. 7132–7141.
40. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2018; pp. 3–19.

41. Zhang, M.; Xu, S.; Song, W.; He, Q.; Wei, Q. Lightweight underwater object detection based on yolo v4 and multi-scale attentional feature fusion. *Remote Sens.* **2021**, *13*, 4706. [CrossRef]
42. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2018; pp. 8759–8768.
43. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2019; pp. 658–666.
44. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 12993–13000. [CrossRef]
45. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: New York, NY, USA, 2017; pp. 2980–2988.
46. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding box regression loss with dynamic focusing mechanism. *arXiv* **2023**, arXiv:2301.10051.
47. Lei, M.; Li, S.; Wu, Y.; Hu, H.; Zhou, Y.; Zheng, X.; Ding, G.; Du, S.; Wu, Z.; Gao, Y. Yolov13: Real-time object detection with hypergraph-enhanced adaptive visual perception. *arXiv* **2025**, arXiv:2506.17733.
48. Dang, C.T.; Sato, H.; Kubo, M. EAW-YOLO11: Enhanced YOLO11 network for underwater object detection. *Artif. Life Robot.* **2025**, *30*, 773–785. [CrossRef]
49. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective kernel networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2019; pp. 510–519.
50. Cheng, T.; Wang, X.; Huang, L.; Liu, W. Boundary-preserving mask r-cnn. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 660–676.
51. Tan, J.; Wang, C.; Li, B.; Li, Q.; Ouyang, W.; Yin, C.; Yan, J. Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2020; pp. 11662–11671.
52. Tan, J.; Lu, X.; Zhang, G.; Yin, C.; Li, Q. Equalization loss v2: A new gradient balance approach for long-tailed object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2021; pp. 1685–1694.