


## Article

# An Improved YOLOv9s Algorithm for Underwater Object Detection

Shize Zhou <sup>1</sup>, Long Wang <sup>1</sup>, Zhuoqun Chen <sup>1</sup>, Hao Zheng <sup>2</sup>, Zhihui Lin <sup>3</sup> and Li He <sup>1,\*</sup> <sup>1</sup> School of Mechatronics and Control Engineering, Shenzhen University, Shenzhen 518000, China; 2310295085@email.szu.edu.cn (S.Z.); 2310295096@email.szu.edu.cn (L.W.)<sup>2</sup> School of Science, Beijing Forestry University, Beijing 100083, China; zhenghaocha2022@163.com<sup>3</sup> School of Medical and Bioinformatics Engineering, Northeastern University, Shenyang 110000, China; 2549880@dundee.ac.uk

\* Correspondence: heli@szu.edu.cn

**Abstract:** Monitoring marine life through underwater object detection technology serves as a primary means of understanding biodiversity and ecosystem health. However, the complex marine environment, poor resolution, color distortion in underwater optical imaging, and limited computational resources all affect the accuracy and efficiency of underwater object detection. To solve these problems, the YOLOv9s-SD underwater target detection algorithm is proposed to improve the detection performance in underwater environments. We combine the inverted residual structure of MobileNetV2 with Simple Attention Module (SimAM) and Squeeze-and-Excitation Attention (SE) to form the Simple Enhancement attention Module (SME) and optimize AConv, improving the sensitivity of the model to object details. Furthermore, we introduce the lightweight DySample operator to optimize feature recovery, enabling better adaptation to the complex characteristics of underwater targets. Finally, we employ Wise-IoU version 3 (WIoU v3) as the loss function to balance the loss weights for targets of different sizes. In comparison with the YOLOv9s model, according to the experiments conducted on the UPRC and Brackish underwater datasets, YOLOv9s-SD achieves an improvement of 1.3% and 1.2% in the mean Average Precision (mAP), reaching 83.0% and 94.3% on the respective datasets and demonstrating better adaptability to intricate underwater environments.

**Keywords:** YOLOv9s; underwater target detection; attention mechanism; upsampling operator; loss function



Academic Editor: Marco Cococcioni

Received: 27 December 2024

Revised: 18 January 2025

Accepted: 22 January 2025

Published: 25 January 2025

**Citation:** Zhou, S.; Wang, L.; Chen, Z.; Zheng, H.; Lin, Z.; He, L. An Improved YOLOv9s Algorithm for Underwater Object Detection. *J. Mar. Sci. Eng.* **2025**, *13*, 230. <https://doi.org/10.3390/jmse13020230>

**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

The development of marine resources is crucial for human economic growth. At the same time, it is essential to effectively protect marine ecosystems during resource development [1]. Monitoring marine organisms enables the observation of changes within marine ecosystems. Underwater target monitoring involves the continuous observation of underwater organisms to gather information about their activity, distribution, and growth characteristics, providing scientific evidence for marine ecosystem protection. However, insufficient underwater lighting and various noise interference significantly degrade image quality, leading to challenges such as low contrast, blurring, and color distortion, which ultimately hinders the accuracy of underwater target monitoring.

In recent years, the rapid advancement of computer vision technology has led to the widespread adoption of deep learning-based methods for detecting underwater biological targets, which have become the dominant approach in this field [2]. Detection methods can

be broadly categorized into two types: two-stage algorithms, represented by SPP-net, Fast R-CNN, and Faster R-CNN [3–5], along with single-stage algorithms, including SSD [6] and the YOLO series [7]. Two-stage algorithms generate candidate regions before detecting targets, offering high accuracy but requiring significant computational resources, which limits their use in resource-constrained settings. In contrast, single-stage algorithms directly predict target classes and bounding boxes, balancing accuracy and speed. However, they may lose features of small targets due to their multi-layer structure. Therefore, the choice of detection algorithm depends on the specific application requirements.

In a single-stage algorithm, the SSD algorithm locates and classifies on the feature map of different scales, while YOLO algorithm predicts the target directly according to the image; it minimizes computational cost and ensures high detection accuracy. Given hardware limitations and real-time detection requirements in underwater environments, YOLO algorithms are widely used [8]. Zhang et al. [9], building on the YOLOv4 network, proposed an Attention Feature Fusion Module (AFFM) that integrates semantic features across various scales, enhancing the recognition of small targets. Li et al. [10] proposed YOLO-TN based on the YOLOv5, applying distillation and pruning to the network structure. They used a specialized network search algorithm to optimize the YOLOv5 backbone, achieving a lightweight model with the 12-fold increase in detection speed. Hou et al. [11] improved the YOLOv5s algorithm by incorporating the HorBlock module within the backbone network to enhance feature extraction capabilities and employing a genetic algorithm for hyperparameter tuning, resulting in improved training accuracy. Chen et al. [12] developed Underwater-YCC based on YOLOv7, embedding the CBAM attention mechanism within a backbone network to improve detection capability and leveraging Conv2Former and Wise-IOU to effectively extract object features, balancing the weighting of both high- and low-quality images. Zhang et al. [13] designed CUIB-YOLO using the YOLOv8n algorithm as its foundation, where the UIB module substitutes the BottleNeck component within the C2f structure to reduce model parameters and incorporates the EMA attention mechanism to enhance the feature processing capabilities. Guo et al. [14] optimized the YOLOv8 network using a lightweight FasterNet backbone to improve computational efficiency, and proposed an FBIFPN structure to solve the lack of target features under multi-scale variations. Cen et al. [15] designed YOLOv9-YX based on YOLOv9, integrating the C3 module and ECA attention mechanism to enhance focus on object features, incorporating the CDown convolution module to reduce computational costs, and proposing the FSPPF multi-scale module to effectively fuse features from different levels.

The underwater environment is complex, with challenges including optical attenuation, scattering, and interference from suspended particles, which result in blurred and smaller targets. These factors place higher demands on target detection technology. As a newer model in the YOLO series, YOLOv9 features a redesigned network architecture based on YOLOv7 [16], significantly improving speed and accuracy. From the perspective of improving accuracy in underwater object detection and lightweight models, this study selects YOLOv9s as the base network. However, the feature extraction layer of YOLOv9s is not sensitive enough to the characteristics of low contrast, blurred edges and color distortion in underwater images. Additionally, the YOLOv9s involves multiple downsampling operations, which can easily lead to the loss of target detail information, and the fixed sampling method during the upsampling stage results in insufficient feature information. This makes it challenging to extract key features effectively, leading to false detection or missed detections. Therefore, it is necessary to adapt YOLOv9s to the unique characteristics of underwater environments to improve its detection capabilities. This study proposes the YOLOv9s-SD detection algorithm. The improvements consist of the following three parts:

1. To suppress background interference, the inverted residual structure concept from MobileNetV2 is leveraged to integrate SimAM and SE attention mechanisms, forming the Simple Enhancement attention Module (SME). The downsampling AConv module is combined with the SME attention mechanism, which enhances the focus of network on object features.
2. To better recover feature information of the target, the DySample upsampling operator is introduced to replace two upsampling operations in YOLOv9s. It adaptively adjusts sampling positions, effectively restoring target edge information, and bypasses dynamic convolution kernels to reorganize feature maps, reducing computational cost.
3. The original CIoU loss function is replaced by WIoU-v3. The WIoU mechanism dynamically adjusts loss weights according to the target size, enabling the adaptation to targets of different sizes and further improving the localization capability for small targets.

Through underwater object detection experiments, YOLOv9s-SD is compared with other object detection algorithms, verifying its effectiveness and facilitating the deployment of underwater object detection models.

## 2. Materials and Methods

### 2.1. YOLOv9 Network Architecture

YOLOv9 [17] ranks among the latest single-stage object detection algorithms, building upon the success of YOLOv7 with significant architectural innovations. The model consists of four main components: the input layer, backbone, neck, and head. The backbone based on the RepNCSPeLan4 module combines the RepConv, Cross Stage Partial (CSP), and Efficient Layer Aggregation Network (ELAN) to optimize feature extraction and gradient flow. RepConv utilizes re-parameterization techniques to enhance inference efficiency, while the CSP and ELAN modules improve feature fusion and the gradient propagation. The neck employs a Feature Pyramid Network (FPN) to fuse multi-scale features, enhancing detection accuracy for objects of varying sizes. Finally, the head adopts an Anchor-Free design, directly predicting bounding box coordinates and class probabilities using a combination of Distribution Focal Loss (DFL) and Complete Intersection over Union (CIoU) loss function.

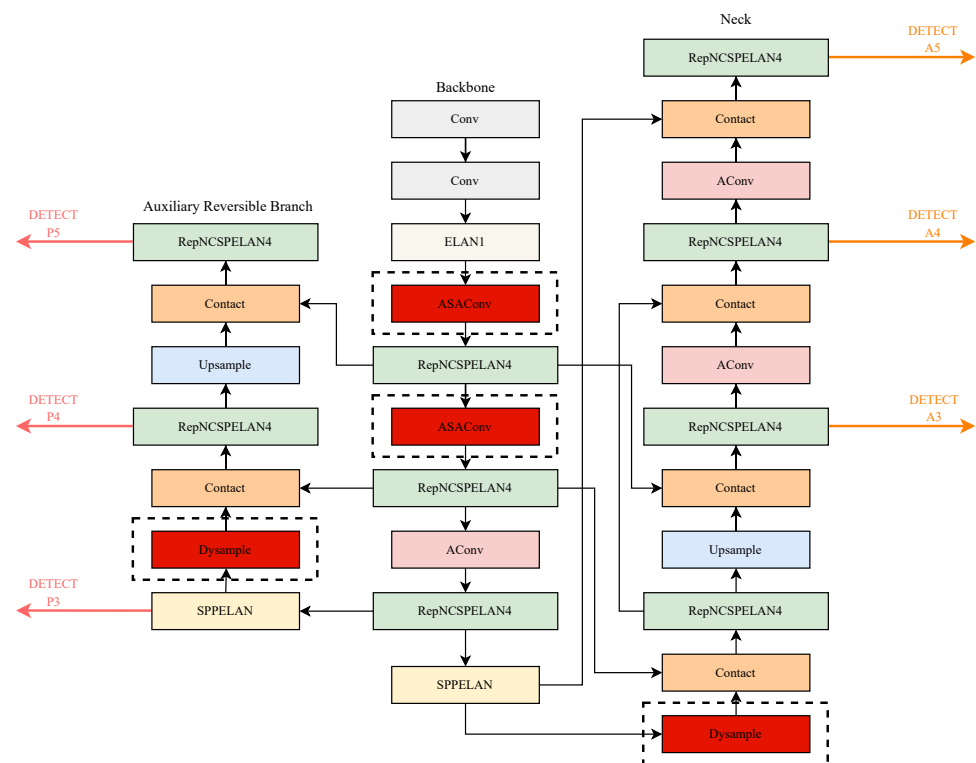
The downsampling module reduces feature map resolution through convolutional and pooling operations while increasing the number of channels, enabling the extraction of higher-level semantic information and expanding the receptive field. By progressively abstracting features from low-level (e.g., edges, textures) to high-level (e.g., shapes, semantics), it provides multi-level feature representations for detection tasks. The detection head uses the CIoU loss function to optimize bounding box predictions. CIoU loss function improves upon traditional IoU by incorporating penalties for center-point distance and aspect ratio, measuring overlap, the distance to center, and the width-to-height ratios. This enhances localization accuracy, particularly for objects of varying sizes or under occlusion.

Two key innovations in YOLOv9 are Programmable Gradient Information (PGI) and the Generalized Efficient Layer Aggregation Network (GELAN). PGI addresses information loss in deep networks through a multi-branch architecture, including a main branch for inference and auxiliary branches that generate reliable gradients during training. GELAN integrates the strengths of CSPNet and ELAN, optimizing parameter utilization and computational efficiency while alleviating information bottlenecks. These advancements enable YOLOv9 to achieve superior detection accuracy and efficiency across various tasks, such as object recognition and image segmentation.

## 2.2. Algorithm Improvements

In the process of underwater target detection, issues such as target blurriness and reduced target image size arise due to underwater optical attenuation, scattering, and interference from suspended particles, leading to false detections and missed detections. Furthermore, it is essential to balance detection accuracy and computational efficiency, improving detection precision while ensuring the model remains lightweight. An enhanced model, YOLOv9s-SD, based on YOLOv9s, is proposed in this paper to solve these challenges.

This paper introduces three key optimizations, as illustrated in Figure 1. First, the SME attention mechanism is embedded in the YOLOv9s backbone, effectively combining spatial and channel feature information to strengthen critical target features and enhance the network's ability to capture details in complex underwater scenarios. Second, the lightweight DySample operator for upsampling is incorporated to substitute certain upsampling modules in the original YOLOv9s network. By dynamically generating upsampling kernels through point sampling, DySample reduces computational overhead, improves adaptability across different feature maps, and optimizes feature recovery capabilities. Finally, the original CIoU loss function is exchanged for the WIoU-v3 loss function, effectively reducing gradient gains between high-quality and low-quality samples, improving bounding box localization precision, and enhancing detection performance for small targets.



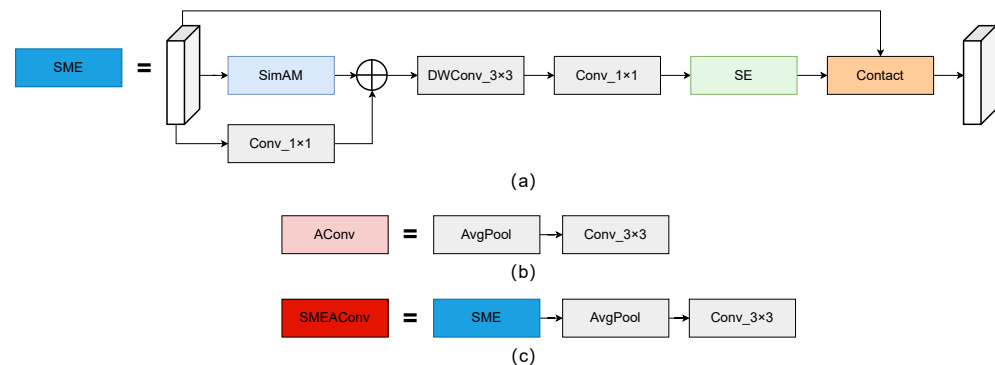
**Figure 1.** The improved YOLOv9s-SD network structure.

### 2.2.1. SME Attention Mechanism

In underwater environments, complex lighting conditions reduce the contrast between targets and backgrounds, making target features unclear. In the YOLOv9s network, downsampling operations are employed to decrease the spatial resolution in feature maps, reducing computational costs as well as memory usage to make the model more lightweight. However, with low-resolution input images, repeated downsampling can easily lead to the loss of target features. In order to improve the ability of underwater targets feature extraction and strengthen how network layers focus on target features, this study integrates the concept of the inverted residual structure from MobileNetV2 [18] with SimAM and



SE attention mechanisms, resulting in the SME attention mechanism, whose structure is shown in Figure 2a. The inverted residual structure enhances nonlinear expression capabilities through a design that first expands and then reduces dimensions, while the residual structure better preserves the information flow between low-level and high-level features.



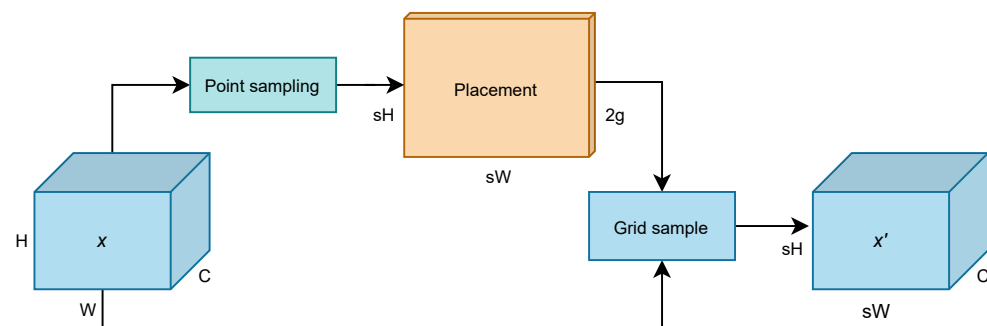
**Figure 2.** The structural framework of each module: (a) structure of the SME attention module; (b) AConv; (c) SMEAConv.

The SME attention mechanism processes feature maps through the SimAM attention module to generate spatial attention features. SimAM [19] calculates the mean and variance of feature maps along spatial dimensions and combines these with activation functions to produce fine-grained spatial weights, highlighting important regions. This can alleviate the impact of low contrast and blurred details in underwater environments. Additionally, a  $1 \times 1$  convolution is used to adjust the number of channels in the input feature map, achieving color compensation and mitigating the impact of color distortion on feature representation. The spatial attention features are then fused with the adjusted features, followed by further feature extraction using depthwise separable convolution (DWConv) and pointwise convolution. This convolution structure enhances feature extraction efficiency while significantly reducing computational overhead. The fused features are further processed by the SE [20] attention module to obtain channel-level weights. SE attention recalibrates channel-wise feature responses, emphasizing the most informative channels while enhancing critical features and suppressing redundant ones. In underwater scenarios, color distortion and uneven lighting conditions often degrade feature representation, and SE attention effectively addresses these issues. Finally, residual connections integrate enhanced features into the original input features, preserving key information from the input and ensuring feature transmission integrity.

In terms of computational cost, the SME attention mechanism is lightweight and efficient, with minimal computational overhead. SimAM is parameter-free, requiring only  $O(H \times W \times C)$  operations for spatial attention. The  $1 \times 1$  convolution and DWConv further reduce computational costs, while the SE module introduces a small number of additional parameters for channel-wise recalibration. Overall, the SME module adds less than 3% additional GFLOPs compared to the baseline YOLOv9s network, making it suitable for resource-constrained underwater object detection tasks. As a result, SME effectively integrates spatial and channel attention features, adaptively capturing spatial location information of targets and strengthening critical features in each channel. The improved structure is presented and illustrated in Figure 2c; it incorporates the SME attention mechanism before the downsampling AConv module, forming the SMEAConv module, which enables the model to focus on key target features while suppressing background noise interference in target localization.

### 2.2.2. DySample Upsampling Operator

In object detection networks, feature upsampling increases feature map resolution to restore and refine the spatial information of targets. Underwater targets are often blurred and small, and YOLOv9s uses nearest-neighbor interpolation for upsampling, resulting in overly smoothed images that struggle to recover details and often lose edge information. Dynamic upsamplers address these challenges by generating content-aware upsampling kernels through dynamic convolution. Examples include methods like CARAFE, FADE, and SAPA [21–23]. However, these methods come with high computational costs. To address the challenges associated with upsampling, this study introduces the DySample [24] upsampling operator. DySample is a lightweight upsampling operator that bypasses dynamic convolution kernels and constructs upsampling through point sampling. Compared to dynamic upsamplers, DySample does not require high-resolution guidance features as input, does not need additional CUDA packages beyond Pytorch, and offers lower inference latency, memory usage, floating-point operations, and parameter count. This approach utilizes feature information more effectively while balancing the trade-off between performance improvement and computational cost, thereby enhancing detection accuracy with minimal overhead. The network structure of the DySample operator is shown in Figure 3.



**Figure 3.** DySample module flowchart.

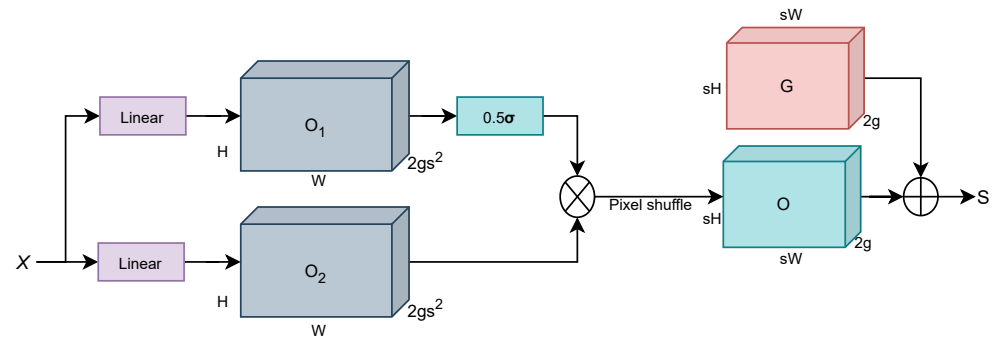
The input feature map  $X$  has a size of  $H \times W \times C$ , and the point sampling set  $S$  has a size of  $sH \times sW \times 2g$ , where  $2g$  represents the x-axis and y-axis coordinates. The Grid sample function is employed to resample the input feature map  $X$  using the point sampling set  $S$ , generating a new feature map  $X'$  of size  $sH \times sW \times C$ , as shown in Equation (1):

$$X' = \text{Grid\_sample}(X, S) \quad (1)$$

The dynamic upsampling operator DySample receives a multi-channel tensor of size  $H \times W \times C$ . It then outputs two offsets,  $O_1$  and  $O_2$ , each of size  $H \times W \times 2gs^2$ , through two parallel linear transformation layers. The sampling factor for  $O_1$  is set to a dynamic factor of  $0.5\delta$ , while the sampling factor for  $O_2$  is set to the default value. To avoid instability caused by excessive offsets, a dynamic range adjustment factor  $O$  is introduced, as shown in Equation (2):

$$O = 0.5 \times \text{sigmoid}(\text{linear}_1(X)) \times \text{linear}_2(X) \quad (2)$$

the offsets  $O_1$  and  $O_2$  are then combined through pixel shuffle and reshaped into an offset  $O$  of size  $sH \times sW \times 2g$ . Finally, the sampling set  $S$  is generated by combining the offset  $O$  with the original data  $G$ , as shown in Equation (3). The process of generating the dynamic point sampling set is illustrated in Figure 4.



**Figure 4.** Dynamic point sampling set generation process.

$$S = G + O \quad (3)$$

Sampling positions are adaptively adjusted by DySample in response to input features; DySample sets the offset ranges, captures critical information, and reorganizes sampled features to enhance upsampling performance. To preserve detailed information of underwater targets and avoid the blurring or aliasing effects caused by traditional pixel-duplication methods, the two upsampling modules in YOLOv9s, marked by red dashed boxes in Figure 1, are replaced with DySample operators, improving target feature recovery and overall detection performance.

### 2.2.3. WIoU-v3 Loss Function

Distant underwater targets occupy fewer pixels, resulting in smaller target areas and making it hard to distinguish these targets from background noise. In order to improve detection accuracy for these targets, selecting an appropriate loss function is crucial. YOLOv9s employs the CIoU [25] loss function to calculate the regression loss of bounding boxes. However, CIoU does not adequately balance complex samples and struggles with the accurate localization of smaller targets. Additionally, the penalty terms added for center-point distance and aspect ratio consistency based on IoU increase computational complexity and overhead. To address these issues, we introduce the WIoU-v3 [26] loss function to replace CIoU. This adjustment directs model focus toward the localization of ordinary-quality boxes while improving attention to the poor-quality boxes. The WIoU-v3 formula is illustrated in Equation (9).

$$L_{IoU} = 1 - IoU \quad (4)$$

$$R_{WIoU} = \exp\left(\frac{(b_{c_x}^{gt} - b_{c_x})^2 + (b_{c_y}^{gt} - b_{c_y})^2}{c_w^2 + c_h^2}\right) \quad (5)$$

where  $b_{c_x}$  and  $b_{c_y}$  represent the center coordinates of the ground truth bounding box,  $b_{c_x}^{gt}$  and  $b_{c_y}^{gt}$  denote the center coordinates of the predicted bounding box, and  $c_w^2$  and  $c_h^2$  represent the length and width of the minimum enclosing rectangle that contains both bounding boxes, respectively.

Consequently, the WIoU-v1 loss function can be expressed as Equation (6):

$$L_{WIoUv1} = R_{WIoU} \times L_{IoU} \quad (6)$$

The quality of the ground truth bounding box can be described by the outlier factor, where the outlier factor  $\beta$  is defined as Equation (7):

$$\beta = \frac{L^*IoU}{L_{IoU}} \in [0, +\infty) \quad (7)$$

where  $L^*IoU$  denotes the focal coefficient. A non-monotonic focusing factor  $r$  based on  $\beta$  can be constructed, and its formula is defined as Equation (8):

$$r = \frac{\beta}{\delta \alpha^{\beta-\delta}} \quad (8)$$

The hyperparameters  $\delta$  and  $\alpha$  are adjusted to suit different models, and the values of  $\alpha$  and  $\delta$  are determined through experimental tuning based on the literature. Here,  $\alpha$  is an empirical value with a default of 1.9, while  $\delta$  is tuned based on the dataset characteristics and model convergence behavior in later stages. A faster convergence rate indicates the effective learning of high-quality samples, and increasing  $\delta$  emphasizes medium- and low-quality samples. In this study, the dataset contains many low-quality samples, and the model converges quickly, as demonstrated in Figure 7. Consequently,  $\delta$  is set to 3. Finally, this non-monotonic focusing factor  $r$  is applied to the WIoU-v1 loss function to obtain WIoU-v3.

$$L_{WIoUv3} = \frac{\beta}{\delta \alpha^{\beta-\delta}} \times L_{WIoUv1} \quad (9)$$

WIoU-v1 was designed as an attention-based prediction box loss, while WIoU-v3 builds upon it by introducing a focal coefficient. This addition reduces the weighting of high-quality samples in the loss through the outlier factor  $\beta$ , dynamically adjusts the gradient gain of bounding boxes, and focuses on medium-quality anchor boxes to strengthen localization performance. Moreover, WIoU-v3 avoids calculations involving aspect ratios, making it more computationally efficient than CIoU. As a result, the WIoU-v3 loss function dynamically optimizes weighting for small targets, significantly enhancing detection accuracy for YOLOv9s-SD.

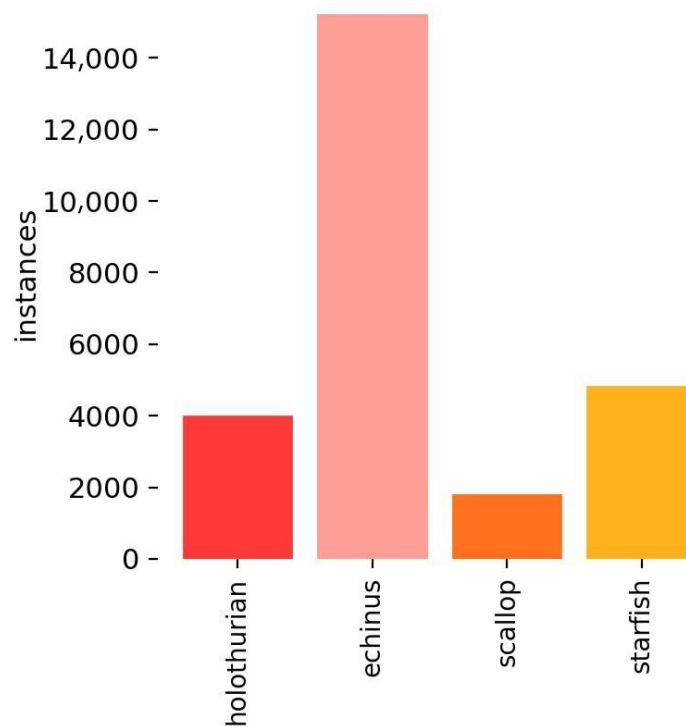
### 3. Experiments and Results

#### 3.1. Experimental Dataset

The dataset used in this study is the URPC [27] dataset, which contains four categories of marine organisms: holothurian, scallop, echinus, and starfish. A total of 6753 images are included; however, due to the effects of light absorption and scattering in underwater environments, the image quality is relatively low. To better simulate real underwater conditions, training the model with low-resolution images helps to enhance the algorithm's generalization ability. Considering the limited computational resources available in underwater applications, using low-resolution images reduces inference computational costs, making the algorithm more lightweight and practical for deployment. Accordingly, the dataset was filtered, and some images with resolutions exceeding  $1920 \times 1280$  were excluded to better reflect real-world conditions, given the blurriness and low contrast of underwater images. Ultimately, 4549 images were retained and split into a training set and a test set in a 7:3 ratio. Table 1 provides statistics on image resolutions. Figure 5 shows the statistical data of biological quantities.

**Table 1.** Dataset resolution statistics.

Resolution/Pixels	Number of Images
586 × 480	44
704 × 576	38
720 × 405	3205
1920 × 1080	645
2048 × 1536	21
2560 × 1440	32
3840 × 2160	2768



**Figure 5.** The statistical data of the quantities of four types of organisms.

### 3.2. Experimental Settings and Evaluation Metrics

All experiments were carried out under the operating system Windows 10 with the following experimental equipment: GPU Nvidia GeForce A5000 (16 GB); and experimental environment: Pytorch1.13.0+Python3.11+CUDA11.3. The experimental parameters were set as follows: input image size of  $640 \times 640$ , an initial learning rate of 0.01, momentum set to 0.937, weight decay coefficient of 0.0005, 300 training epochs, and a batch size of 16. For underwater biological detection, six metrics were employed to accurately evaluate the model performance, including Precision ( $P$ ), Recall ( $R$ ), Average Precision ( $AP$ ), Mean Average Precision ( $mAP$ ), Giga Floating Point Operations (GFLOPs), and model size. The relevant calculation formulas are as follows.

$$P = \frac{TP}{TP + FP} \quad (10)$$

$$R = \frac{TP}{TP + FN} \quad (11)$$

$$AP = \sum_{i=0}^1 P(R_i) \cdot \Delta R_i \quad (12)$$

$$mAP = \frac{1}{n} \sum_{i=1}^n AP_i \quad (13)$$

In this formula,  $TP$  represents the quantity of positive samples correctly identified,  $FP$  represents the quantity of negative samples misclassified into the positive category, and  $FN$  indicates the quantity of positive samples incorrectly classified into the negative category.

### 3.3. Analysis of Experimental Results

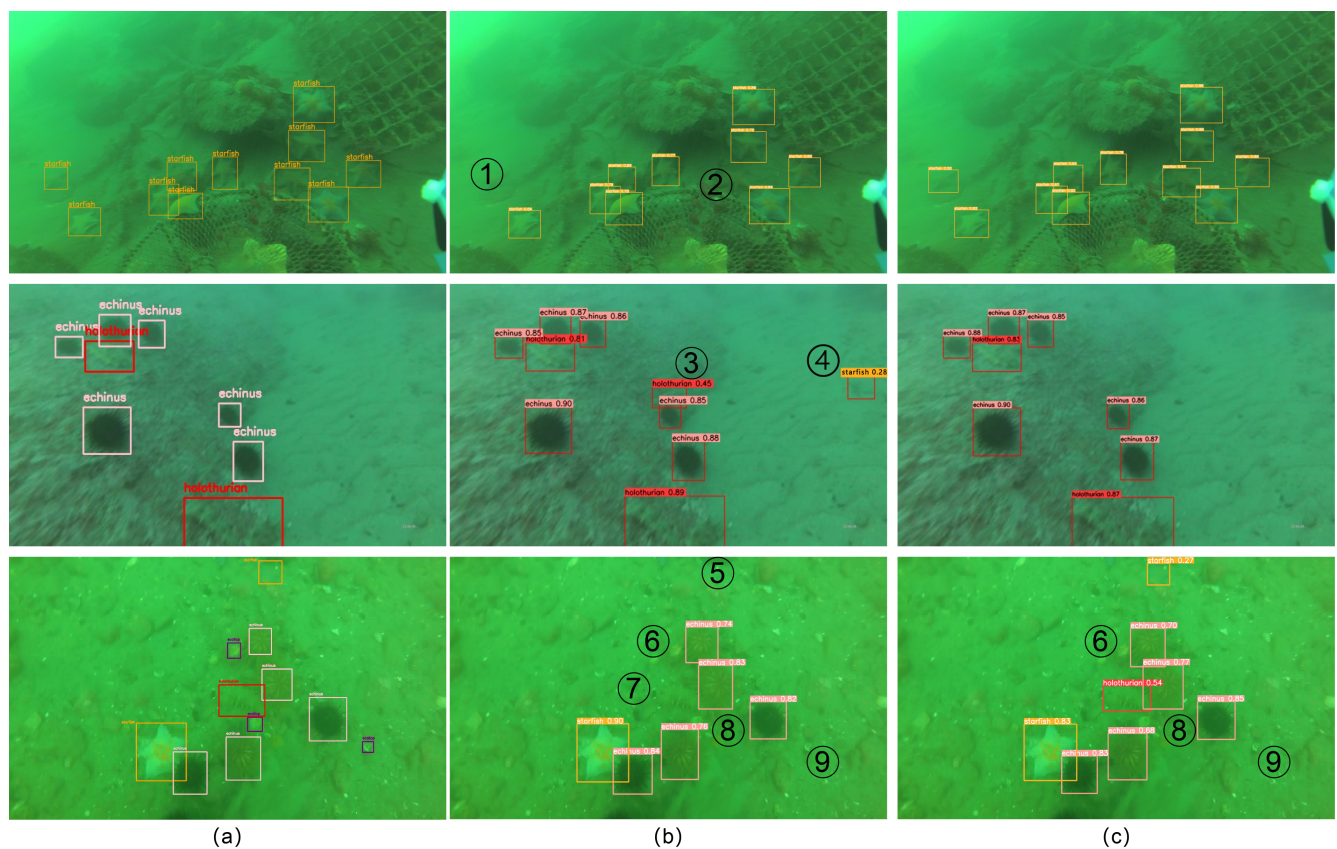
To validate the efficacy of the improved algorithm, the baseline YOLOv9s along with the improved YOLOv9s-SD models were trained on the same dataset. Table 2 presents the final experimental results.



**Table 2.** Comparative results of YOLOv9s and YOLOv9s-SD.

Model	P/%	R/%	mAP50 (%)	mAP50:95 (%)	GFLOPs	Size/MB
YOLOv9s	81.9	75.6	81.7	50.6	39.6	14.3
YOLOv9s-SD	83.6	76.7	83.0	52.1	40.5	14.4

The experimental data indicate that the YOLOv9s-SD model achieves an increase of 1.7% in precision, 1.1% in recall, 1.3% in mAP50%, and 1.5% in mAP50:95%, with only a 2.27% increase in computational cost. These results demonstrate that the YOLOv9s-SD model improves detection accuracy for underwater targets, particularly in capturing targets in low-resolution images. Furthermore, the model size increases by just 0.7%, maintaining its lightweight nature and meeting the deployment requirements for underwater target detection. Detailed experimental results are presented in Figure 6.



**Figure 6.** Experimental detection results before and after YOLOv9s improvement; (a) real label; (b) YOLOv9s detection results; (c) YOLOv9s-SD detection results.

The detection results in Figure 6b show that the original model missed detections for starfish at positions ①, ②, and ⑤, and falsely detected a holothurian at position ③ and a starfish at position ④. Additionally, the model did not detect the holothurian at position ⑦. By incorporating the self-developed SME attention mechanism module, the improved model effectively reduces false detections and enhances detection capability in scenarios where the target and background are similar. However, both the YOLOv9s-SD model and the YOLOv9s model failed to detect scallops at positions ⑥, ⑧, and ⑨, indicating that the detection performance for smaller targets, such as scallops, still requires further improvement.

### 3.4. Attention Mechanism Comparison Experiment

For the purpose of verifying the effectiveness of different attention mechanisms, common attention mechanisms including CA [28], SE, ECA [29], SimAM, CBAM [30], MLCA [31], and SCSA [32] were individually added to the same location for comparative experiments. The experimental findings are detailed in Table 3.

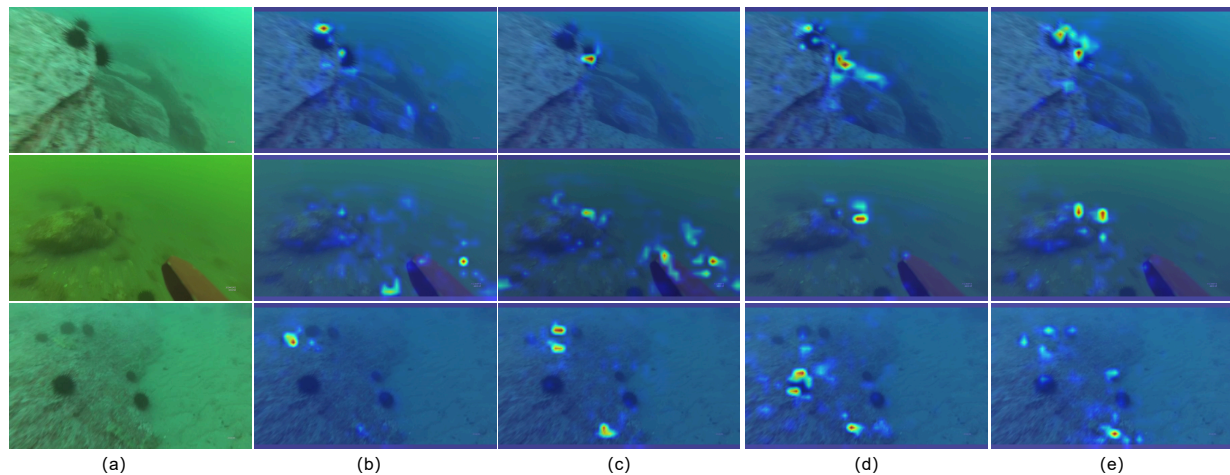
**Table 3.** Comparison experiment of different attention mechanisms.

Model	mAP50 (%)	mAP50:95 (%)	GFLOPs	Size/MB
YOLOv9s	81.7	50.6	39.6	14.3
+CA	82.0	50.8	39.6	14.3
+SE	82.0	50.8	39.6	14.3
+ECA	81.9	50.8	39.6	14.3
+SimAM	82.1	50.6	39.6	14.3
+CBAM	81.7	50.7	39.6	14.3
+SCSA	81.9	50.8	39.6	14.3
+MLCA	82.0	50.9	39.6	14.3
+SME	82.3	51.1	40.5	14.4

Based on the data presented in Table 3, the addition of various attention mechanisms improves the accuracy of underwater target detection with minimal changes to model size and computational cost. CBAM, which combines channel and spatial attention, did not improve mAP50%, likely due to its difficulty in handling the low contrast and complex backgrounds typical of underwater environments. ECA increased the mAP50% by 0.2% by enhancing channel-wise features but lacks spatial modeling, limiting its effectiveness for small and low-contrast targets. SCSA achieved a 0.2% increase in mAP50% by combining spatial and channel-wise attention and leveraging the collaborative potential of multi-semantic information, while its performance is constrained by the noisy and low-visibility conditions of underwater scenes. CA, SE, and MLCA achieved a 0.3% increase in mAP50%, with CA and SE focusing on channel-wise relationships and MLCA capturing cross-layer dependencies, making them effective for multi-scale targets but less suited to the dynamic nature of underwater environments. SimAM further improved mAP50% by 0.1% through its parameter-free design, offering a computationally efficient solution, while its simplicity limits its ability to solve the diverse and complex features present in underwater scenes. Finally, SME outperformed all others, increasing mAP50% by 0.6% and mAP50:95% by 0.5%. SME combines the strengths of SimAM and SE, making it particularly effective for detecting small and low-contrast targets in underwater scenarios. This comparison highlights the superiority of SME in underwater target detection.

To better visualize the enhanced feature extraction capabilities of the SME attention mechanism, the fifth layer within the network employed the LayerCAM [33] algorithm, and the results were visualized using heatmaps. The visualization results are shown in Figure 7.

The visualization results indicate that the added SME attention mechanism more effectively integrates spatial and channel features of the image, enhances feature extraction, and suppresses background interference in target localization. Consequently, incorporating the SME attention mechanism markedly improves detection accuracy.



**Figure 7.** Displaying heatmap visualization results of different attention mechanisms using LayerCAM algorithm: (a) raw image; (b) YOLOv9s; (c) YOLOv9s+SE; (d) YOLOv9s+SimAM; (e) YOLOv9s+SME.

### 3.5. Ablation Experiment

To validate the effectiveness of every modified module, YOLOv9s served as the baseline model for the ablation experiments. Detection performance was analyzed on the URPC dataset, and the ablation experiment results are illustrated in Table 4. In the table, ✓ signifies the introduction of the module, whereas × signifies its exclusion.

From the results in Table 4, it can be observed that adding the SME attention mechanism improved mAP50% by 0.6%. The Average Precision (AP) for each target increased, with starfish and scallop showing AP improvements of 0.6% and 0.9%, respectively. This enhancement improved the model's capability to extract target features from complex backgrounds.

**Table 4.** Ablation study.

Experiments	SME	DySample	WOUv3	mAP50 (%)	Holothurian (%)	Starfish (%)	Scallop (%)	Echinus (%)
1	×	×	×	81.7	80.4	88.1	66.3	92.0
2	✓	×	×	82.3	80.9	88.7	67.2	92.4
3	✓	✓	×	82.6	81.6	87.9	68.1	92.7
4	✓	✓	✓	83.0	81.8	88.5	69.2	92.6

Introducing the DySample upsampling module further increased mAP50% by 0.3%. The AP for holothurian increased by 1.2%, scallop by 1.8%, and echinus by 0.7%, reaching its best performance. The network's capability to restore target feature information was strengthened. However, the AP for starfish slightly decreased due to the DySample upsampling module's inability to effectively restore its edge contour information. Furthermore, replacing the CIoU loss function with WIoU-v3 resulted in an additional mAP50% improvement of 0.4%, further optimizing the localization performance of bounding boxes. While echinus experienced a slight decrease in AP, the AP for other targets increased. In the dataset, Figure 5 shows that echinus is the most abundant and has distinct color features, making it easier to identify. The WIoU-v3 loss function focuses on balancing the weights of some low-quality samples, prioritizing their anchor boxes, which contributed to the decrease in AP for echinus.

When all modules were incorporated into the model, its overall performance reached its best, with mAP50% improving by 1.3% compared to the original model. Notably, the

model improved significantly detection accuracy for small targets and those with similar backgrounds, with the AP of holothurians and scallops increasing by 1.4% and 2.9%, respectively.

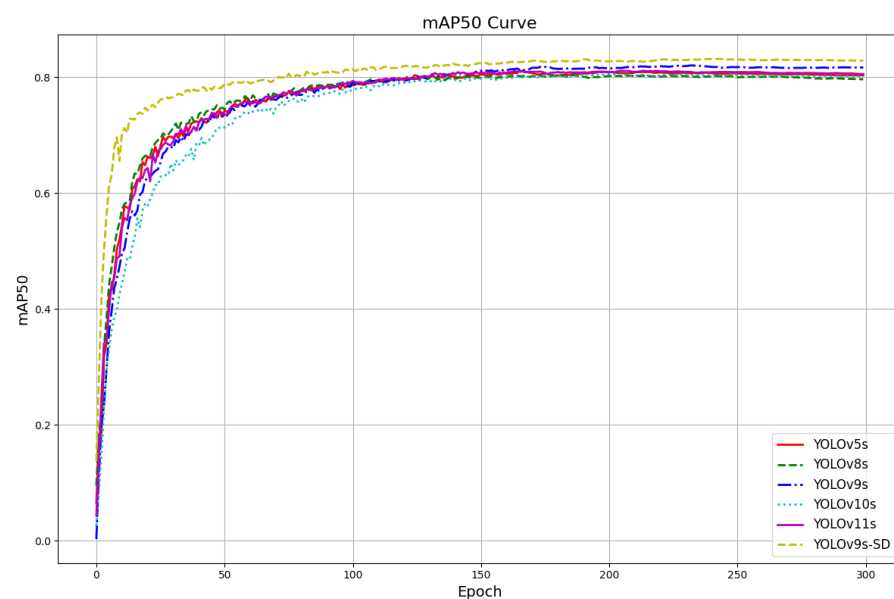
### 3.6. Comparison Experiment

To further validate the detection performance of the YOLOv9s-SD model, several widely used object detection algorithms were selected for comparison, including YOLOv5s, YOLOv7-tiny, YOLOv8s, YOLOv10s [34], and YOLOv11s [35]. The training and testing of all the models were conducted under the same experimental conditions. Detailed comparison results can be found in Table 5 and Figure 8.

**Table 5.** Comparison experiment.

Model	P/%	R/%	mAP50 (%)	mAP50:95 (%)	GFLOPs	Size/MB
YOLOv5s	81.5	74.8	80.5	48.7	19.0	16.0
YOLOv7-tiny	81.3	74.2	78.7	44.5	13.2	12.3
YOLOv8s	83.2	73.2	80.6	50.0	23.6	19.0
YOLOv10s	83.1	74.6	80.4	50.2	24.8	15.7
YOLOv11s	81.9	75.6	81.3	49.3	21.6	19.2
YOLOv9s	81.9	75.6	81.7	50.6	39.6	14.3
YOLOv9s-SD	83.6	76.7	83.0	52.1	40.5	14.4

The experimental results indicate that YOLOv9s-SD demonstrates superiority across multiple performance metrics compared to other models. YOLOv9s-SD achieved a precision (P) of 83.6% and a recall (R) of 76.7%, showing improvements in both metrics, which indicates a higher number of correctly identified samples and excellent performance. Its mAP50% and mAP50:95% reached 83.0% and 52.1%, respectively, outperforming other models. Additionally, while maintaining high accuracy, YOLOv9s-SD has a size of 14.4 MB and a computational cost of 40.5 GFLOPs. In comparison with traditional models such as YOLOv5s and YOLOv8s, the smaller size of the YOLOv9s-SD model enhances its adaptability for lightweight deployment in underwater target detection tasks. Therefore, under conditions of limited underwater computational resources, the proposed improved model meets the requirements for accuracy and a lightweight design, making it suitable for underwater environments.

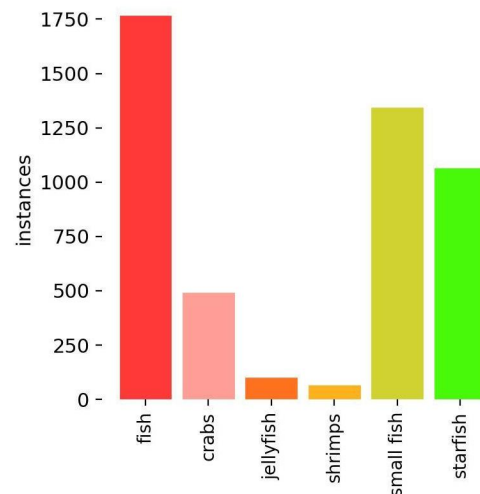


**Figure 8.** The mAP50% curve for each model.



### 3.7. Generalization Experiments

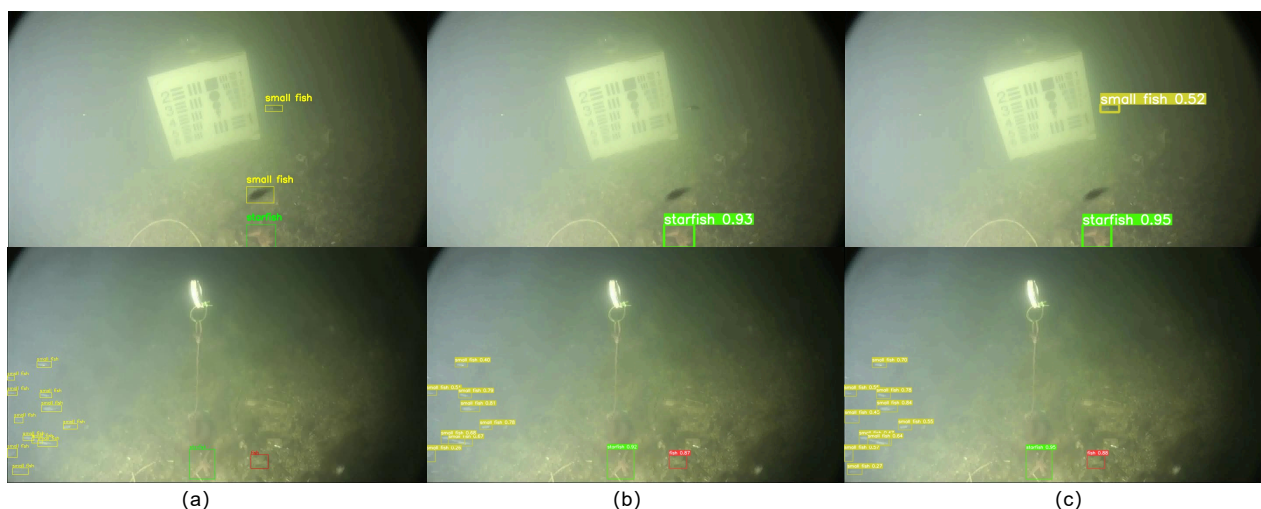
To further validate the generalization performance of the YOLOv9s-SD model, experiments were conducted using the Brackish [36] dataset. Organisms are classified into six categories: fish, small fish, crabs, shrimps, jellyfish, and starfish. The dataset contains a total of 2465 images, with low resolution and a uniform size of  $960 \times 540$ . Additionally, the statistical distribution of the number of organisms is shown in Figure 9. The dataset was split into a training set and a test set in a 7:3 ratio. All models were trained and tested under the same experimental conditions as in Section 3.2. The detailed experimental outcomes are displayed in Table 6 and Figure 10.



**Figure 9.** The statistical data of the quantities of six types of organisms.

**Table 6.** Comparison experiment.

Model	P/%	R/%	mAP50 (%)	mAP50:95 (%)	GFLOPs	Size/MB
YOLOv5s	96.3	87.3	94.2	70.8	19.0	17.6
YOLOv7-tiny	90.4	78.9	84.7	52.9	13.2	11.7
YOLOv8s	94.1	89.1	93.3	71.4	23.6	21.4
YOLOv10s	95.1	85.8	92.8	70.3	24.8	15.7
YOLOv11s	92.3	90.8	90.4	69.7	21.6	18.2
YOLOv9s	92.7	89.0	93.1	70.0	39.6	14.3
YOLOv9s-SD	94.9	89.0	94.3	72.2	40.5	14.4



**Figure 10.** Experimental detection results before and after YOLOv9s improvement; (a) real label; (b) YOLOv9s detection results; (c) YOLOv9s-SD detection results.



The experimental results indicate that YOLOv9s-SD achieved the best overall performance. Compared to the YOLOv9s model, it achieved a 1.2% improvement in mAP50% and a 2.2% increase in mAP50:95%. Additionally, compared to the YOLOv5s and YOLOv8s models, its mAP50% improved by 0.1% and 1%, respectively. In Figure 10, it is shown that the YOLOv9s model has a problem in that it sometimes fails to detect targets, while the improved YOLOv9s-SD model can detect more small fish targets. Therefore, the improved model satisfies the demands of underwater target detection tasks.

#### 4. Discussion

The detection experiment results on the two datasets reveal that the YOLOv9s-SD algorithm still has certain shortcomings. On the URPC dataset, targets such as scallops, which appear smaller due to imaging factors, result in missed detections. This indicates that the detection performance of the YOLOv9s-SD algorithm for smaller targets needs further improvement. Additionally, although the self-developed attention mechanism has enhanced the algorithm's ability to extract target features, while more effective targets were detected on the Brackish dataset, not all targets were successfully identified, and the precision of the model still requires further enhancement. Most of the missed targets are smaller in size, with indistinct features that are difficult to differentiate from the background. To enhance the ability of the model to effectively extract these features and improving detection accuracy for such targets. Future research can leverage the imaging characteristics of underwater environments to effectively utilize color information [37] for image enhancement. By applying image enhancement techniques to preprocess underwater images, the quality of the images can be improved, better highlighting the detailed features of the targets. This approach will benefit underwater object detection tasks.

#### 5. Conclusions

To address the complexity of underwater scenarios and the challenges posed by blurred and small targets, which limit the feature extraction capabilities of target detection algorithms, we propose the YOLOv9s-SD algorithm. The SME attention mechanism is integrated into the AConv downsampling module to enhance target feature extraction capabilities. A lightweight DySample upsampling operator is introduced to restore target details, and the WIoU-v3 loss function is employed to improve localization accuracy for small targets. Validation on the URPC and Brackish datasets shows that the proposed algorithm increases mAP50% by 1.3% and 1.2%, respectively. This demonstrates its ability to be deployed in resource-constrained underwater environments and effectively achieve target detection in such scenarios. The proposed method has significant potential applications in marine ecological monitoring, underwater resource exploration, and autonomous underwater vehicle (AUV) navigation. By providing an efficient and accurate target detection solution, this research contributes to the advancement of underwater robotics and supports sustainable marine ecosystems. Despite its promising performance, YOLOv9s-SD has certain limitations. For example, detecting extremely small targets in highly noisy environments remains challenging.

Considering the characteristics of underwater environments, future research can focus on optimizing the model's structure and designing more efficient attention mechanisms. Advanced data augmentation techniques can be explored to enhance the robustness and accuracy of algorithms in underwater settings. Furthermore, integrating domain adaptation methods to improve generalization across diverse underwater environments could be a valuable direction.

**Author Contributions:** Conceptualization, S.Z.; methodology, S.Z. and L.W.; software, S.Z. and Z.C.; validation, S.Z. and Z.L.; formal analysis, Z.L. and H.Z.; writing—original draft preparation, S.Z.; writing—review and editing, S.Z. and L.H.; visualization, S.Z. and H.Z.; supervision, L.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by National key research and development plan (2022YFB3904602), Shenzhen Basic Research Key Project (JCYJ20220818095816035).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data and results supporting the findings of this study can be obtained from the corresponding author upon reasonable request.

**Acknowledgments:** The author thanks the members of the laboratory for their help and cooperation during the experiment.

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

## References

1. Greenville, J.; MacAulay, T. Protected areas in fisheries: A two-patch, two-species model. *Aust. J. Agric. Resour. Econ.* **2006**, *50*, 207–226. [\[CrossRef\]](#)
2. 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. [\[CrossRef\]](#)
3. 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. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
5. 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.* **2017**, *39*, 1137–1149. [\[CrossRef\]](#) [\[PubMed\]](#)
6. 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 Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
7. Jiang, P.; Ergu, D.; Liu, F.; Cai, Y.; Ma, B. A review of YOLO algorithm developments. *Procedia Comput. Sci.* **2022**, *199*, 1066–1073. [\[CrossRef\]](#)
8. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
9. Zhang, M.; Xu, S.; Song, W.; He, Q.; Wei, Q. Lightweight underwater object detection based on YOLOv4 and multi-scale attentional feature fusion. *Remote. Sens.* **2021**, *13*, 4706. [\[CrossRef\]](#)
10. Li, W.; Li, Y.; Li, R.; Shen, H.; Li, W.; Yue, K. Research on Rapid Detection of Underwater Targets Based on Global Differential Model Compression. *J. Mar. Sci. Eng.* **2024**, *12*, 1760. [\[CrossRef\]](#)
11. Hou, C.; Guan, Z.; Guo, Z.; Zhou, S.; Lin, M. An improved YOLOv5s-based scheme for target detection in a complex underwater environment. *J. Mar. Sci. Eng.* **2023**, *11*, 1041. [\[CrossRef\]](#)
12. Chen, X.; Yuan, M.; Yang, Q.; Yao, H.; Wang, H. Underwater-YCC: Underwater target detection optimization algorithm based on YOLOv7. *J. Mar. Sci. Eng.* **2023**, *11*, 995. [\[CrossRef\]](#)
13. Zhang, Q.; Chen, S. Research on improved lightweight fish detection algorithm based on YOLOv8n. *J. Mar. Sci. Eng.* **2024**, *12*, 1726. [\[CrossRef\]](#)
14. Guo, A.; Sun, K.; Zhang, Z. A lightweight YOLOv8 integrating FasterNet for real-time underwater object detection. *J. Real-Time Image Process.* **2024**, *21*, 49. [\[CrossRef\]](#)
15. Cen, Q.; Zhu, Q.; Wang, Y.; Chen, W.; Liu, S. YOLOv9-YX: Lightweight algorithm for underwater target detection. *Vis. Comput.* **2024**, 1–13. [\[CrossRef\]](#)
16. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
17. Wang, C.Y.; Yeh, I.H.; Mark Liao, H.Y. YOLOv9: Learning what you want to learn using programmable gradient information. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September 29–4 October 2024; pp. 1–21. [\[CrossRef\]](#)

18. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
19. Yang, L.; Zhang, R.; 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; PMLR: New York, NY, USA, 2021; pp. 11863–11874.
20. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
21. Wang, J.; Chen, K.; Xu, R.; Liu, Z.; Loy, C.C.; Lin, D. Carafe: Content-aware reassembly of features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3007–3016.
22. Lu, H.; Liu, W.; Fu, H.; Cao, Z. FADE: Fusing the assets of decoder and encoder for task-agnostic upsampling. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 231–247.
23. Lu, H.; Liu, W.; Ye, Z.; Fu, H.; Liu, Y.; Cao, Z. SAPA: Similarity-aware point affiliation for feature upsampling. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 20889–20901.
24. Liu, W.; Lu, H.; Fu, H.; Cao, Z. Learning to upsample by learning to sample. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 6027–6037.
25. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000. [[CrossRef](#)]
26. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding box regression loss with dynamic focusing mechanism. *arXiv* **2023**, arXiv:2301.10051.
27. Sahoo, A.; Dwivedy, S.K.; Robi, P.S. Advancements in the field of autonomous underwater vehicle. *Ocean. Eng.* **2019**, *181*, 145–160. [[CrossRef](#)]
28. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
29. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
30. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.
31. Wan, D.; Lu, R.; Shen, S.; Xu, T.; Lang, X.; Ren, Z. Mixed local channel attention for object detection. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106442. [[CrossRef](#)]
32. Si, Y.; Xu, H.; Zhu, X.; Zhang, W.; Dong, Y.; Chen, Y.; Li, H. SCSA: Exploring the synergistic effects between spatial and channel attention. *arXiv* **2024**, arXiv: 2407. 05128.
33. Jiang, P.T.; Zhang, C.B.; Hou, Q.; Cheng, M.M.; Wei, Y. LayerCAM: Exploring hierarchical class activation maps for localization. *IEEE Trans. Image Process.* **2021**, *30*, 5875–5888. [[CrossRef](#)]
34. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. Yolov10: Real-time end-to-end object detection. *arXiv* **2024**, arXiv:2405.14458.
35. Khanam, R.; Hussain, M. Yolov11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725.
36. Pedersen, M.; Brulund Haurum, J.; Gade, R.; Moeslund, T.B. Detection of marine animals in a new underwater dataset with varying visibility. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–20 June 2019; pp. 18–26.
37. Garcia-Mateos, G.; Hernandez-Hernandez, J.; Escarabajal-Henarejos, D.; Jaen-Terrones, S.; Molina-Martinez, J. Study and comparison of color models for automatic image analysis in irrigation management applications. *Agric. Water Manag.* **2015**, *151*, 158–166. [[CrossRef](#)]

**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.