


Article

Underwater Robot Target Detection Algorithm Based on YOLOv8

Guangwu Song , Wei Chen *, Qilong Zhou and Chenkai Guo

College of Automation, Jiangsu University of Science and Technology, Zhenjiang 212003, China; 221210301116@stu.just.edu.cn (G.S.); 221210301130@stu.just.edu.cn (Q.Z.); 221110303108@stu.just.edu.cn (C.G.)
* Correspondence: cw1@just.edu.cn

Abstract: Although the ocean is rich in energy and covers a vast portion of the planet, the present results of underwater target identification are not sufficient because of the complexity of the underwater environment. An enhanced technique based on YOLOv8 is proposed to solve the problems of low identification accuracy and low picture quality in the target detection of current underwater robots. Firstly, considering the issue of model parameters, only the convolution of the ninth layer is modified, and the deformable convolution is designed to be adaptive. Certain parts of the original convolution are replaced with DCN v3, in order to address the issue of the deformation of underwater photos with fewer parameters and more effectively capture the deformation and fine details of underwater objects. Second, the ability to recognize multi-scale targets is improved by employing SPPFCSPC, and the ability to express features is improved by combining high-level semantic features with low-level shallow features. Lastly, using WIoU loss v3 instead of the CIoU loss function improves the overall performance of the model. The enhanced algorithm mAP achieves 86.5%, an increase of 2.1% over the YOLOv8s model, according to the results of the testing of the underwater robot grasping. This meets the real-time detection needs of underwater robots and significantly enhances the performance of the object detection model.

Keywords: object detection; YOLOv8; convolutional neural networks; ROV underwater robot



Citation: Song, G.; Chen, W.; Zhou, Q.; Guo, C. Underwater Robot Target Detection Algorithm Based on YOLOv8. *Electronics* **2024**, *13*, 3374. <https://doi.org/10.3390/electronics13173374>

Academic Editor: Cecilio Angulo

Received: 6 July 2024

Revised: 10 August 2024

Accepted: 21 August 2024

Published: 25 August 2024



Copyright: © 2024 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 ocean accounts for the vast majority of the Earth's total surface area and contains abundant oil, natural gas, minerals, chemicals, and aquatic resources [1]. However, due to the presence of more influential factors in the underwater environment, the detection results are often unsatisfactory when performing underwater target detection [2]. In recent years, the rapid development of deep learning technology, especially the introduction of convolutional neural networks, has brought new possibilities and applications for object detection in images [3,4]. Deep learning technology can efficiently and accurately complete difficult tasks like object recognition, picture segmentation, and image classification by using deep neural network models. Therefore, there is substantial research value in the field of marine science for using deep learning technology for quick and accurate underwater target detection.

Deep-learning-based object detection methods can be divided into two-stage algorithms and one-stage algorithms [5]. The two-stage algorithm first identifies the candidate regions containing the target, and then classifies and locates the target, mainly using the Fast R-CNN (Fast Region-based Convolutional Neural Network) [6] algorithm and the Faster R-CNN (Fast Region-based Convolutional Neural Network) [7] algorithm. Its modest detection speed and great detection accuracy are its hallmarks. The one-stage algorithm does not need to generate candidate boxes containing targets, but directly classifies and locates the targets, mainly using the SSD (Single Shot MultiBox Detector) algorithm [8], the YOLO (You Only Look Once) series algorithm [9], and the Resnet (Residual Network) [10], which

offers the one-stage algorithm good detection accuracy and excellent detection speed. This makes it appropriate for real-time detection jobs, since it makes direct predictions about categorization and localization. Because of the limitations of underwater gear, single-stage object detection methods are currently used for most underwater target detection.

Underwater target detection is obviously more difficult in practical applications because of the diversity and complexity of underwater settings, which lead to the typically inferior image quality and target clarity obtained by underwater robots. Wei et al. [11] improved the YOLOv3 model by adding compression and excitation modules, as well as expanding the detection scale. Zhang et al. [12] proposed a lightweight underwater object detection method that combines MobileNet v2 and YOLOv4 algorithms with attention feature fusion to achieve a balance between accuracy and speed in object detection in marine environments. Li et al. [13] achieved a fast and accurate underwater marine organism detection method by incorporating attention mechanism and multi-scale detection strategy into the improved YOLOv5 model, combined with image enhancement and optimized prediction head structure. Lei et al. [14] proposed an improved YOLOv5 object detection algorithm suitable for complex underwater environments by using Swin Transformer as the basic backbone network of YOLOv5, improving the path aggregation network for multi-scale feature fusion and optimizing the confidence loss function. Li et al. [15] proposed an improved YOLOv5s real-time fish target detection network, which replaces the original backbone network with a ShuffleNetv2 lightweight network using the SE channel attention mechanism, and uses an improved BiFPN short network for feature fusion, achieving model lightweighting and improved detection accuracy. Zhang et al. [16] proposed an improved YOLOv5 underwater target detection network, which improves accuracy and reduces missed detections by adding a global attention mechanism, introducing a DAMO-YOLO-based neck fusion module, and using SIoU (Scale-Invariant IoU) loss. Based on the experimental data, the model performs better.

Li et al. [17] used the YOLOv7 model to improve the accuracy of small target detection by enhancing feature retention and reducing feature loss, introducing spatial pyramid pooling and cross-level partial channel modules, as well as integrating coordinate attention modules. The experimental results show that the improved model is superior to the original model and other recent algorithms in reducing error and improving average accuracy. Liu et al. [18] proposed an improved YOLOv7 network, which replaces the original structure with an ACmixBlock module, integrates jump connections and a 1×1 convolutional architecture, designs a ResNet ACmix module, inserts a global attention mechanism, and optimizes anchor boxes using the K-means++ algorithm, thereby improving the accuracy of feature extraction and network inference speed. Chen et al. [19] proposed an underwater YCC optimization algorithm based on YOLOv7, which integrates the convolutional block attention module (CBAM) to capture fine-grained semantic information. In addition, Conv2Former is used as a network neck component to handle underwater blurred images. Finally, the WIoU (weighted Intersection over Union) loss method was used to effectively improve the detection accuracy. Although these schemes have made progress in certain aspects, they do not involve a fully consideration of the transferability of the model and do not support fast detection. For object identification technology to be applied in marine environments, it needs to have lighter model characteristics and greater real-time processing capabilities. Therefore, creating a lightweight underwater target identification method with a small model volume and high detection accuracy is especially crucial and urgent.

This article presents an upgraded YOLOv8s network model that uses the YOLOv8 model from the YOLO series as its foundation, taking into account the background mentioned above. The model is lightweight, has enhanced detection accuracy, and maintains a sufficient detection speed to meet the needs of multi-scale underwater target detection. The main contributions of this article are as follows:

- (1) Due to the frequent occurrence of image deformation, if the convolution process is still executed according to the preset fixed path, the processing efficiency will be greatly affected. For this reason, we only adjust the convolutional layer in layer 9, and

replace the original convolutional method with adaptive deformable convolutional DCN v3, which can capture changes in and subtle features of underwater targets more efficiently and effectively deal with the challenge of the deformation of the underwater image with fewer parameters.

- (2) The utilization of SPPFCSPC (spatial pyramid pooling-fast with cross-stage partial connections) to improve multi-scale target recognition and to improve feature expression through the integration of shallow, low-level characteristics with high-level semantic features.
- (3) When WIoU loss v3 is substituted for the CIoU (complete intersection over union) loss function, the model’s overall performance improves.

In order to test the effectiveness of the improved model, we applied it to underwater robots and made improvements to the YOLOv8 model in the three previously mentioned areas. We also compared the improved model to other popular algorithms, and the results demonstrate that optimal outcomes were obtained.

2. Underwater Target Detection Algorithm Based on Improved YOLOv8

2.1. YOLOv8 Object Detection Algorithm

The YOLOv8 algorithm is part of the YOLO family of algorithms, which builds on the success of its predecessors by integrating innovative features that significantly improve performance and applicability. In the realm of computer vision, this approach is extensively employed for tasks including object tracking, instance segmentation, picture classification, and object recognition, among others. As shown in Figure 1, the YOLOv8 structure is primarily composed of three parts: the Backbone, Neck, and Head. To meet different performance and resource requirements, YOLOv8 offers five different scale models: YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x.

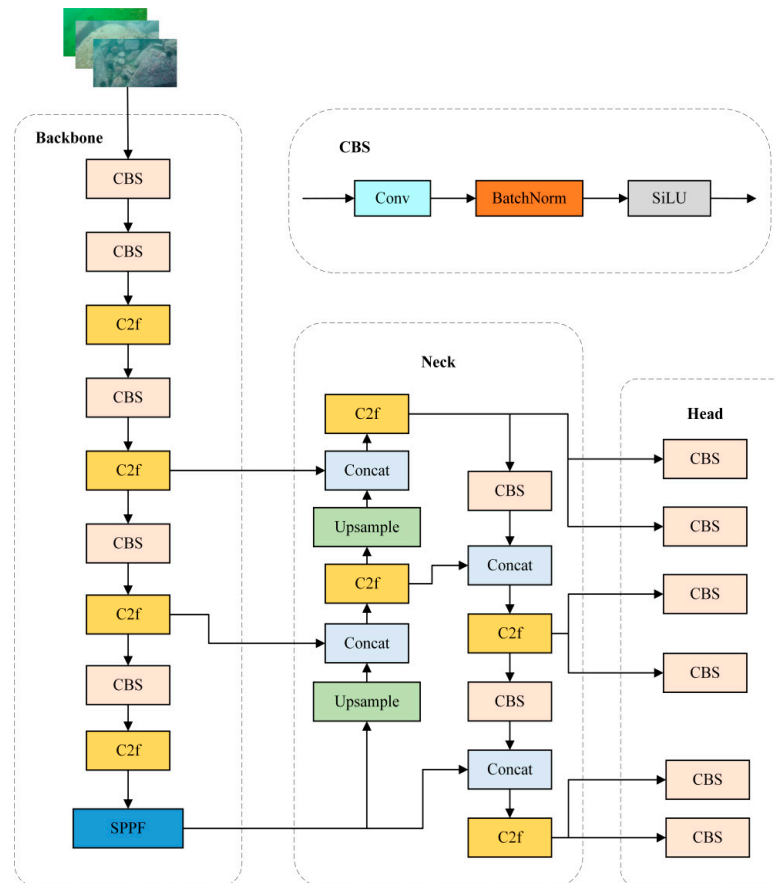


Figure 1. YOLOv8 network structure diagram.

Firstly, the Backbone segment uses the framework of CSPDarknet53 (cross-stage partial network Darknet53) [20], which includes basic convolutional units (Conv) and spatial pyramid pooling fast (SPPF) for local and global feature fusion. The SPPF module uses pooling and convolution techniques to adaptively mix feature data of different sizes in order to enhance the model’s ability to extract features. Additionally, the C2f module is incorporated into CSPDarknet53, enhancing the gradient flow of the model and enhancing feature extraction capacity by broadening the depth and receptive field of the network. Figure 2 illustrates the C2f module’s precise structure.

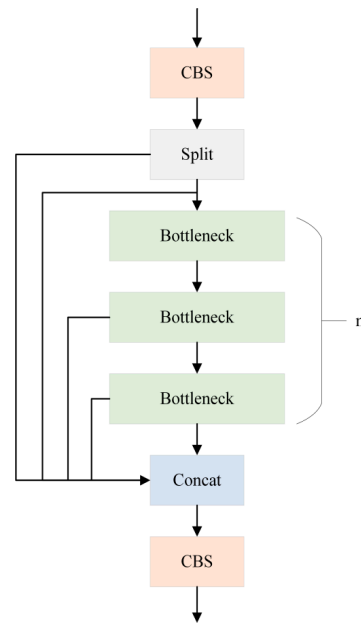


Figure 2. C2f module structure.

C2f input is usually the feature map or raw image output from the previous layer. Firstly, a convolution operation is performed to change the dimension and size of the input features. Following a sequence of convolution and pooling operations, the output is the feature map, which is then sent to the following layer for processing. This procedure reduces the size of the feature map while maintaining the relevant information.

Secondly, the Neck section adopts the PAN FPN (path aggregation network and feature pyramid networks) structure, which fuses multi-scale feature maps by processing the features taken from Backbone. The C2f module is repurposed in this framework to improve feature processing capabilities.

Finally, the Head section separates classification and detection using the current mainstream Decoupled Head structure, thereby reducing the conflict between localization and classification; concurrently, the object detection process adopts the Anchor Free mechanism, which performs better on targets of irregular length and width. Figure 3 depicts the decoupling head model structure.

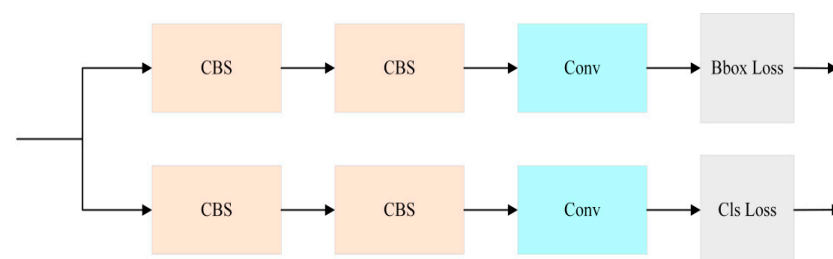


Figure 3. Decoupling head model structure.

Following receipt of the feature maps from the feature pyramid network, the decoupling head divides and processes the classification and regression tasks in the detection task independently; the target’s class is predicted using the classification loss (Cls loss), and its bounding box location and size are predicted using the regression loss (Bbox loss).

2.2. Improvement of YOLOv8 Object Detection Network

2.2.1. Improvement of Original Convolution

In traditional convolutional networks [21], the pooling layer keeps the downsampling ratio constant while the convolutional kernel changes the picture along a predetermined route and step size. Image distortion is more common in special environments, such as underwater. The adaptive DCNv3 (Deformable ConvNet v3) [22] effectively solves the problem of image distortion in underwater target recognition. Unlike traditional convolution, it introduces additional offset variables into the convolution kernel, allowing the kernel to flexibly change shape and position based on image content, thereby more accurately capturing the features of deformed images.

The following is the conventional convolution calculation formula:

$$y(p_0) = \sum_{k=1}^x w_k x(p_0 + p_k) \tag{1}$$

where p_0 represents the center position, k signifies the quantity of sample points, p_k represents the k -th position of the sampling grid, and w_k shows the appropriate sampling sites’ projection weight. Deformable convolution adds an offset matrix Δp_k on the basis of standard convolution, which transforms the convolution into irregular convolution. The following is the deformable convolution calculation formula:

$$y(p_0) = \sum_{k=1}^K w_k x(p_0 + p_k + \Delta p_k) \tag{2}$$

where Δp_k represents the offset. Figure 4 compares the convolution kernel with added offset to the standard convolution kernel. Figure 4a shows the regular convolution, while Figure 4b shows the deformable convolution.

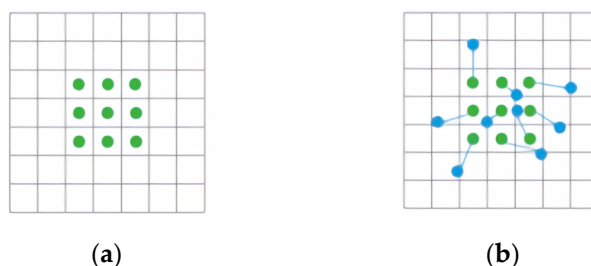


Figure 4. A comparative analysis of enhanced offset convolution and conventional convolution. (a) The standard convolution; (b) the deformable convolution.

In this study, we used DCN v3 instead of the original convolution. DCN v3 introduces a multi grouping mechanism based on DCN v2 [23], whereby each group may improve the capacity of aquatic species to represent their features through factor modulation, sampling vector projection, and independent offset sampling. Sampling vector projection’s primary purpose is to transform the continuous offset into a discrete grid coordinate offset, while factor modulation dynamically adjusts the size and direction of the offset during the learning process, flexibly adapting to the diversity of target deformation at different positions and scales, and optimizing the feature extraction effect. In addition, DCN v3 further introduces a normalization mechanism for modulation scalars, effectively solving the instability problems that DCN v2 may encounter during training, greatly enhancing the model’s performance and stability.

The calculation formula for DCN v3 can be represented by Equation (3):

$$y(p_0) = \sum_{g=1}^G \sum_{k=1}^K w_g m_{gk} x_g(p_0 + p_k + \Delta p_{gk}) \tag{3}$$

where G symbolizes the overall quantity of aggregation clusters, K indicates how many sample points there are for the g -th category, w_g represents the projection weight of the group, m_{gk} represents the group's K -th sample point's modulation scalar, x_g represents the segmented feature map, and Δp_{gk} represents the sampling offset that matches the sampling position p_k in the g -th group.

In order to create C2f-DCN v3, this post integrates the C2f module in YOLOv8 with a deformable convolutional network. The module structure is shown in Figure 5. This can improve the model's target identification performance even further, particularly when handling objects with intricate shape changes and geometric deformations. Furthermore, as shown in Figure 6, by altering the convolution of the Bottleneck portion in C2f, the convolution kernel is able to adaptively shift its spatial location, better capturing the target's form variations.

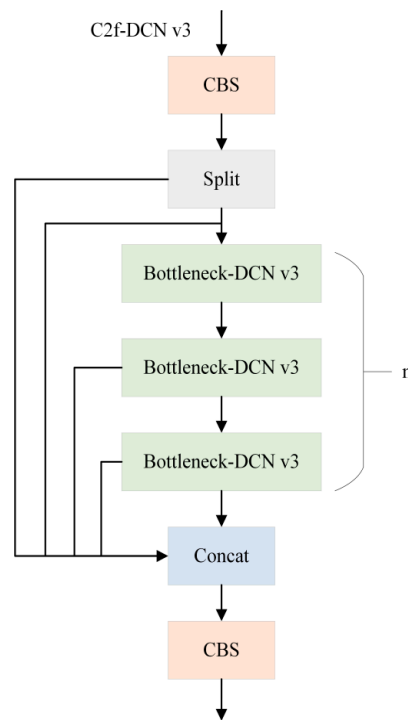


Figure 5. C2f-DCN v3 structure.

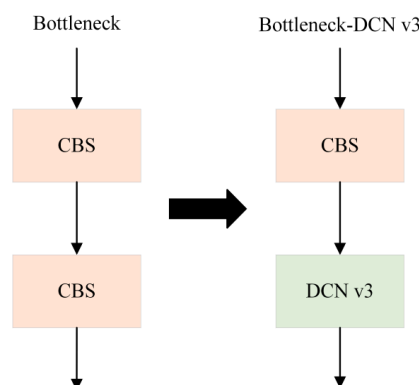


Figure 6. Bottleneck improvements, before and after.

2.2.2. Improvement of Spatial Pyramid Pooling

The input feature map is serially passed through several 5×5 max pooling layers using SPPF in order to extract and fuse high-level features, fuse local and global features, and output an adaptive size. The SPPF structure is shown in Figure 7.

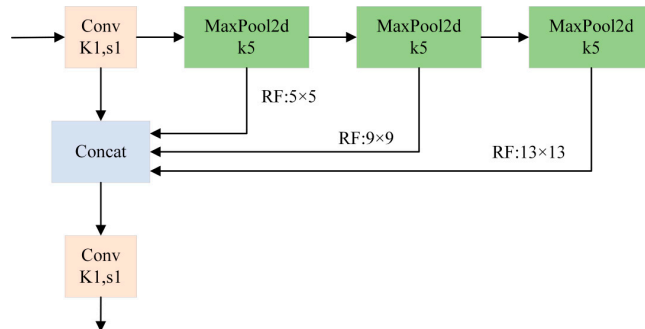


Figure 7. SPPF structure diagram.

Figure 8 illustrates the construction of SPPCSPC, which combines the two approaches of spatial pyramid pooling (SPP) and cross-stage partial connection (CSPC). Typically, the SPP module has a pyramid structure with several layers, each of which corresponds to a distinct pooling scale. The feature map is split up into varying numbers of blocks at each level, and each block is pooled. By capturing data in various sizes, the network may enhance its feature extraction skills. The CSPC module separates the features into two subsets, one of which is processed using ordinary convolution and the other using spatial pyramid pooling. These two sections’ characteristics are combined at the end.

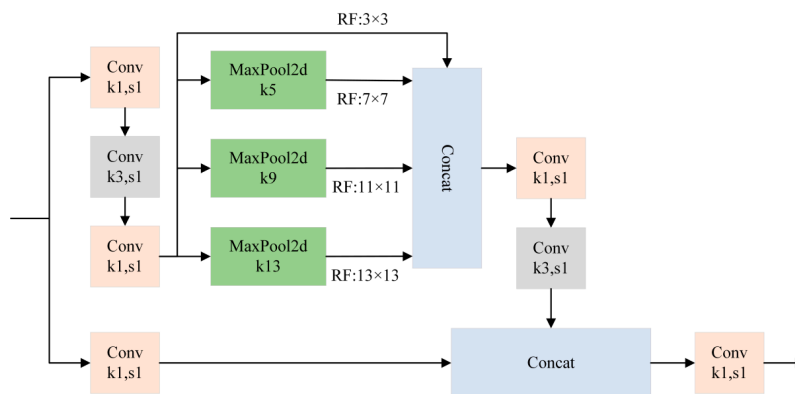


Figure 8. SPPCSPC structure diagram.

Conventional convolution only processes some features, while the SPP structure processes all features, avoiding redundant calculations. This can effectively promote the interaction and fusion of features at different stages, maintain high accuracy, reduce computational complexity, and improve the real-time processing speed and capability of underwater robots.

In order to avoid image distortion, duplicate feature extraction, and improve the algorithm’s feature fusion ability, in this study, we adopted the SPPFCSPC (spatial pyramid pooling-fast and cross-stage partial connection) structure to replace SPPF, and Figure 9 depicts its network structure. Large underwater targets require the model to be able to record a wider range of spatial information, which is made possible by SPPFCSPC, which broadens the network’s receptive field. In an underwater environment, the size of a target can vary significantly due to the distance between the target and the camera. Target size variations may be better accommodated by the detection model thanks to SPPFCSPC’s capacity to encode feature information at different scales. Furthermore, SPPFCSPC main-

tains good performance while optimizing the network structure, lowering computational complexity, and accelerating the model’s runtime.

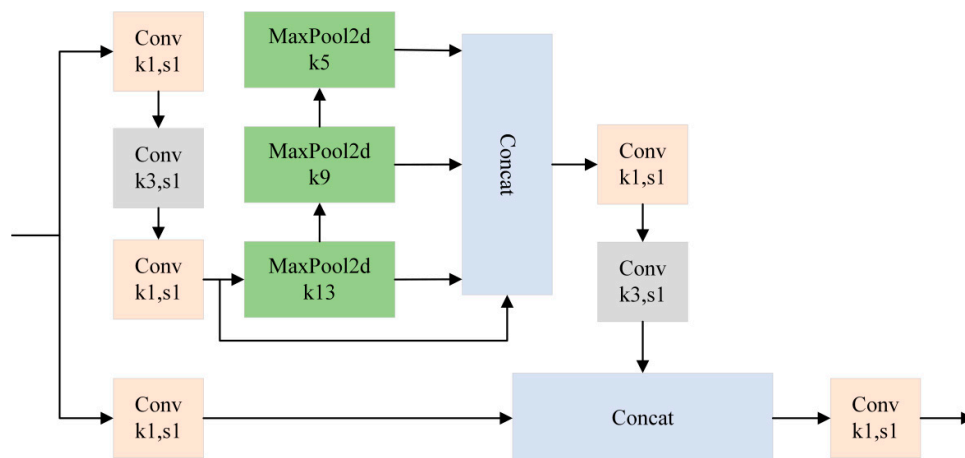


Figure 9. Structure diagram of SPPFCSPC.

2.2.3. Improvement of Loss Function

The regression branch and the classification branch are the two halves of the loss function in YOLOv8, where the regression branch uses distribution focal loss [24] and CIoU loss [25]. While intersection over union ratio (IoU) is an indicator for assessing object recognition accuracy and is used to evaluate the coverage between the predicted box and the genuine box, regression loss is used to investigate the divergence between the predicted box and the actual box. The following is its formula for calculation:

$$IoU = \frac{A \cap B}{A \cup B} \tag{4}$$

The actual bounding box and the expected bounding box are displayed in A and B , respectively. Figure 10 displays the schematic diagram of the IoU parameters.

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \tag{5}$$

where b and b^{gt} show the centers of the expected box and the real box, respectively, $\rho^2(b, b^{gt})$ shows the distance in Euclides between the center points of the predicted and actual boxes, c shows the length of the diagonal of the smallest bounding box between the actual and anticipated boxes, α represents a positive equilibrium parameter, and v demonstrates how the expected and actual boxes’ aspect ratios match. The following is the calculation formula for α and v :

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{6}$$

$$v = \frac{4}{\pi^2} \left(\arctan \frac{w^{gt}}{h^{gt}} - \arctan \frac{w}{h} \right)^2 \tag{7}$$

where w^{gt} , h^{gt} , w , and h respectively illustrate the variations in height and width between the expected and real boxes.

Predicted box assessment accuracy is greatly increased by CIoU loss, which takes into account the overlap area, center point distance, and aspect ratio of the real and predicted boxes from various viewpoints. This allows for a more precise assessment of the relative positions of the two boxes.

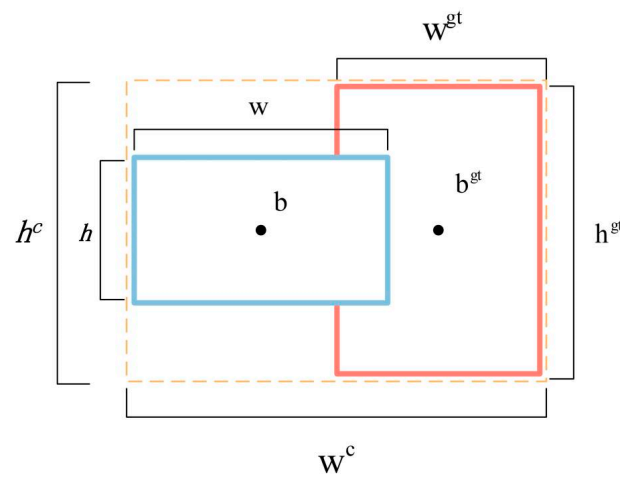


Figure 10. Schematic diagram of IoU parameters.

However, when there is a linear relationship between the projected and real boxes' aspect ratios, the aspect ratio penalty term cannot work as intended. This might have an impact on the predicted box's regression. WIoU is a bounding box regression loss function used in target detection. This loss function is characterized in that it employs a dynamic non-monotonic focusing mechanism to improve the performance of bounding box regression when dealing with samples of varying quality. An intelligent method of assigning gradient enhancement is put into place, and the quality of the anchoring frame is assessed using an outlier metric. This means alleviating the competitive pressure on high-quality anchored frames while reducing the negative gradient impact from poor-quality samples. In light of this, in this article, we suggest using a novel loss function called WIoU loss v3 [26] in instead of CIoU loss in order to balance the effects of varying picture quality on model training and enhance the precision of detection findings. The following is the formula used to calculate WIoU loss:

$$L_{WIoUv1} = R_{WIoU} L_{IoU} \tag{8}$$

$$R_{WIoU} = \exp\left(\frac{\rho^2(b, b^{gt})}{c^2}\right) \tag{9}$$

where the weight decrease of standard anchor boxes was R_{WIoU} enhanced, although the reduction in weight of superior anchor boxes was L_{IoU} reduced R_{WIoU} .

In accordance with $WIoUv1$, $WIoUv3$ constructs non-monotonic focusing coefficients r through an outlier β , and the following is the computation formula:

$$\beta = \frac{L_{IoU}^*}{L_{IoU}} \in [0, +\infty) \tag{10}$$

$$r = \frac{\beta}{\delta \cdot \alpha^{\beta-\delta}} \tag{11}$$

$$L_{WIoUv3} = r \cdot L_{WIoUv1} \tag{12}$$

where β represents outlier, reflecting the quality of the regression box, α , and δ is a hyper-parameter. The regression box can achieve the maximum gradient gain when its outlier, β , is equal to the predetermined value. The acquired image quality may differ due to the intricate undersea environment. WIoU v3's dynamic non-monotonic focusing mechanism, which can dynamically alter the loss function's focusing point in accordance with the quality of the anchor frames, allows it to adapt to a variety of datasets and difficult settings. While IoU is static and uses the same processing for all anchor frames, WIoU v3

can process samples with different qualities more efficiently by evaluating the degree of anomalies of the anchor frames instead of relying only on the IoU values. Additionally, the gradient gain assignment strategy of WIoU v3 assigns gradients intelligently to lessen the harmful gradients generated by low-quality samples and to lessen the competitiveness of high-quality anchor frames. This helps the model to concentrate more on the samples that are more important for performance improvement.

2.3. Improved YOLOv8 Object Detection Network

In response to the actual situation of ROV, firstly, due to the perspective effect of underwater scenes and the influence of target distance, DCN v3 is adopted in the model to more effectively capture the shape deformation and subtle features of underwater targets. Secondly, considering the complexity of model parameters, only the convolutional layer of the ninth layer was adjusted, and SPPFCSPC was introduced to enhance the recognition ability of multi-scale targets. Ultimately, to mitigate the possible influence of subpar underwater picture quality on model performance, a unique loss function, called WIoU, was implemented in place of the traditional CIoU loss function. This enhanced the feature expression by combining low-level textural qualities with high-level semantic information. Figure 11 depicts the network architecture of the YOLOv8s model following these focused enhancements.

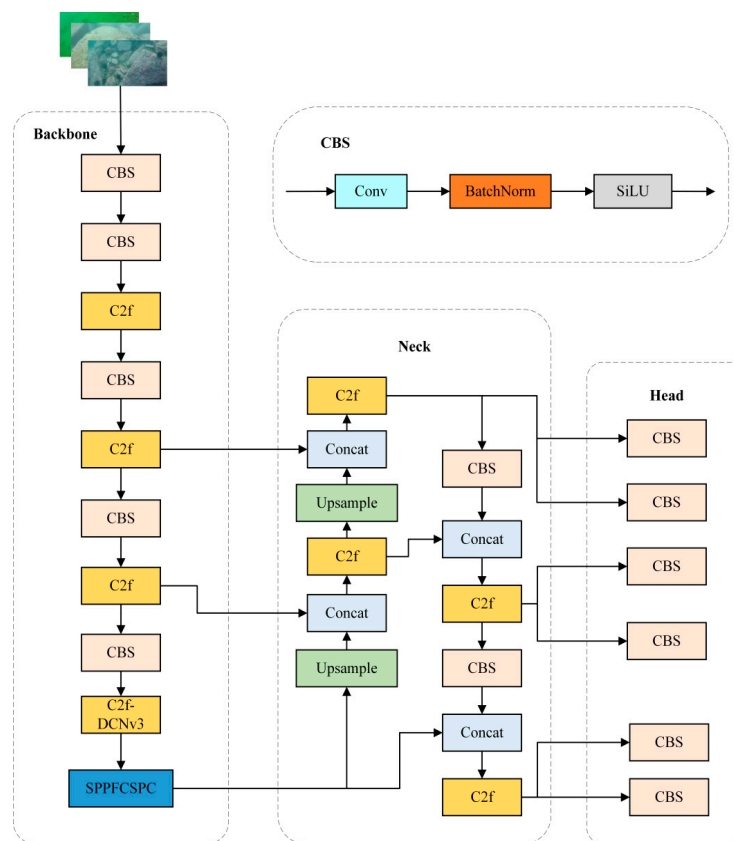


Figure 11. Improved YOLOv8 network structure diagram.

3. Experiments and Analysis

3.1. Datasets and Experimental Environment

The URPC (Underwater Robot Professional Challenge) dataset is used in this paper. As shown in Figure 12, in this study, we compiled photographs of echinus, holothurian, scallops, and starfish from the network as an enlarged dataset, totaling 5543 underwater target images. The URPC dataset has been published by the Underwater Robotics Professional Competition Organization since 2017, with the aim of promoting research into and the

development of underwater vision technology for use in research areas such as underwater target detection, recognition, and classification. The collection contains photographs of a variety of underwater environments, such as muddy seas and clean seawater.

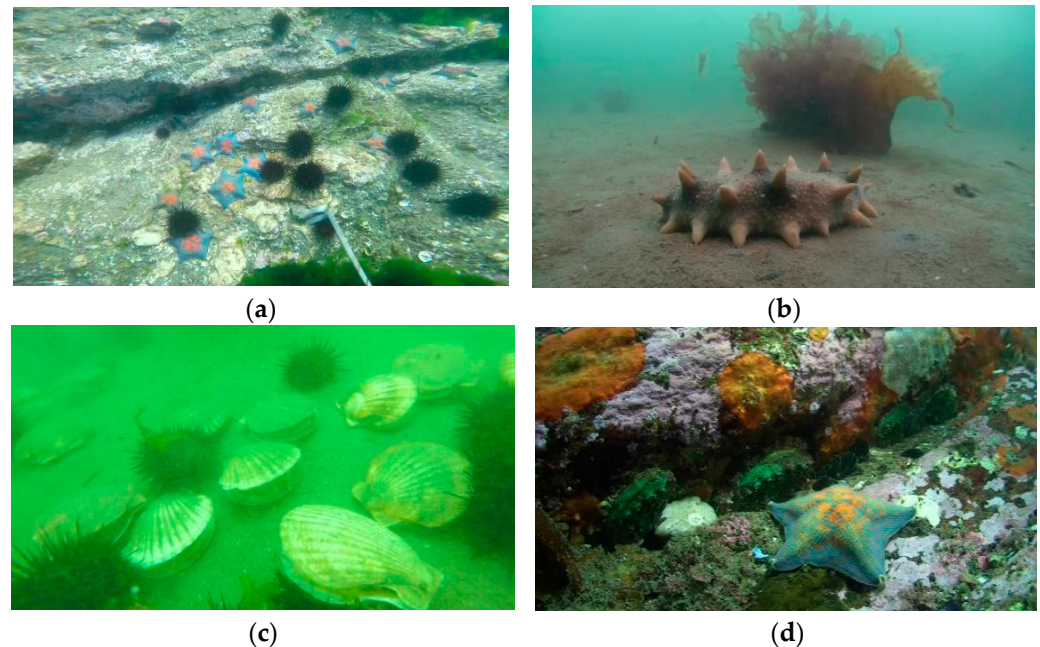


Figure 12. Dataset images. (a) Echinus; (b) holothurian; (c) scallop; (d) starfish.

This dataset contains multiple types of objects, as shown in Figure 13a. Among them, sea urchins have the highest number, followed by starfish, scallops, and sea cucumbers. The bounding box size distribution is shown in Figure 13b, where the majority of the items have sizes that are concentrated in the range of 0.0 to 0.2. In order to promote sufficient learning of the designed model on the dataset, a training set and a validation set comprised 9:1 of the dataset, respectively.

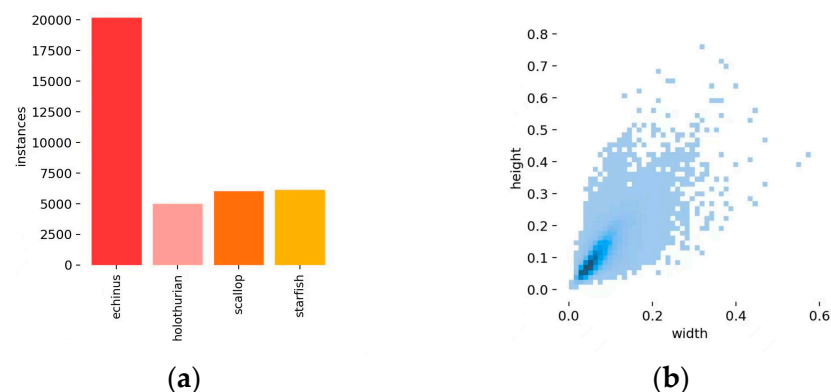


Figure 13. Data Distribution. (a) Histogram of dataset classification; (b) the distribution of bounding box dimensions.

The following is the experimental setup used for this article: GPU, NVIDIA GeForce GTX 3060; CPU, Intel (R) Core (TM) i9-12900; memory, 16.00 GB; graphics memory, 6 GB. The development environment is PyCharm, the programming language is Python, and the operating system and software environment are Windows11 + CUDA11.7 + Python 3.10 + pytorch 1.12.1. Table 1 shows the fundamental parameter settings.

Table 1. Fundamental parameter configurations.

Parameter	Value
Batch size	4
Learning rate	0.01
Optimizer	SGD
Weight attention factor	0.0005
Confidence threshold	0.5

3.2. Underwater Robot Experimental Platform

This experimental platform uses a 20-kg-class ROV (remotely operated vehicle) underwater robotic system, which mainly consists of an aquatic console and an underwater motion unit. The above-water console is equipped with a NVIDIA 3080 high-performance image processor to ensure efficient and accurate image processing. The underwater motion unit is equipped with a 1080p high-definition low-light camera, capable of clear imaging within a range of 3 m, and is directly connected to the industrial control computer. The underwater robot system is shown in Figure 14.

**Figure 14.** Physical image of ROV underwater robot.

3.3. Evaluation Indicators

This paper provides a comprehensive evaluation of the underwater target-identification performance of object detection algorithms using frames per second (FPS), mean average precision (mAP), and average precision (AP).

Precision, sometimes called detection precision, is the number of samples with positive predictions represented as a percentage of the actual number of positive samples. The formula is as follows:

$$Precision = \frac{TP}{TP + FP} \quad (13)$$

Recall (R) is the proportion of correctly identified positive samples. The following is the formula:

$$Recall = \frac{TP}{TP + FN} \quad (14)$$

A higher score denotes better model performance. The $F1$ score is the harmonic mean of recall and accuracy. A high $F1$ score indicates good recall and accuracy performance of the model. An $F1$ score may be computed using the formula below:

$$F_1 = \frac{2 * precision * recall}{precision + recall} = \frac{2TP}{2TP + FP + FN} \quad (15)$$

For every undersea category, AP stands for average probability of correct prediction. The formula is as follows:

$$AP = \int_0^1 PdR \tag{16}$$

where R stands for recall. The average and sum of the AP indicators for each category is known as mAP . This is a complete indicator that considers both Precision and Recall. The model’s accuracy is often assessed using $mAP@0.5$, where the model’s accuracy is set to 0.5 and the average of all categories is computed. The following is the formula:

$$mAP = \frac{\sum_{i=1}^N AP_i}{N} \tag{17}$$

where N symbolizes the entire quantity of categories. The algorithm’s speed in detecting objects is indicated by the number of frames broadcast per second, or FPS.

3.4. Experimental Results and Analysis

3.4.1. Verification of the Enhanced YOLOv8 Object Detection Algorithm

Figure 15 displays the Precision–Recall curve of the model before and after the improvement, and it can be seen from the figure that the $mAP@50$ of the improved model reached 86.5%, an increase of 2.2% compared with that before the improvement, among which the color of the sea urchin was obvious, the AP value of the echinus was 92.1%, and the AP value of holothurian was 76.5% due to its color and background. Figure 16 displays the F1 value curve of the model before and after the improvement, and it can be seen that the improved model also increased compared with the original model.

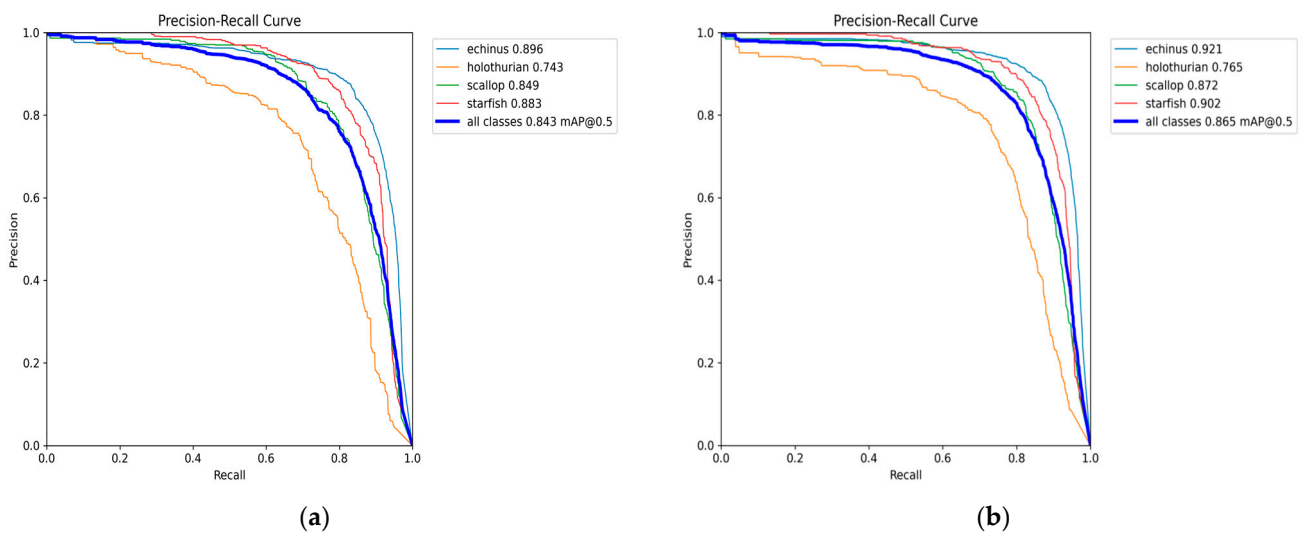


Figure 15. Precision–Recall curve. (a) The before effect; (b) the after effect.

Figure 17 displays the parameter curves of the new model, the mAP value of the model, which is 2.1% higher than the prior model, at 86.5%, and the loss curves for the training and validation sets. The findings demonstrate that the enhanced model outperforms the original YOLO v8 model in terms of detecting effect.

Figure 18 displays the prediction graph for the validation set of the enhanced model. Underwater target identification accuracy is significantly increased by the enhanced model’s capacity to identify unlabeled targets through analysis and comparison with the labeled dataset. This ability is a reflection of the model’s adaptability and network feature extraction capability. When compared to the old model, it is evident that the upgraded model performs better in the target identification of submerged objects.

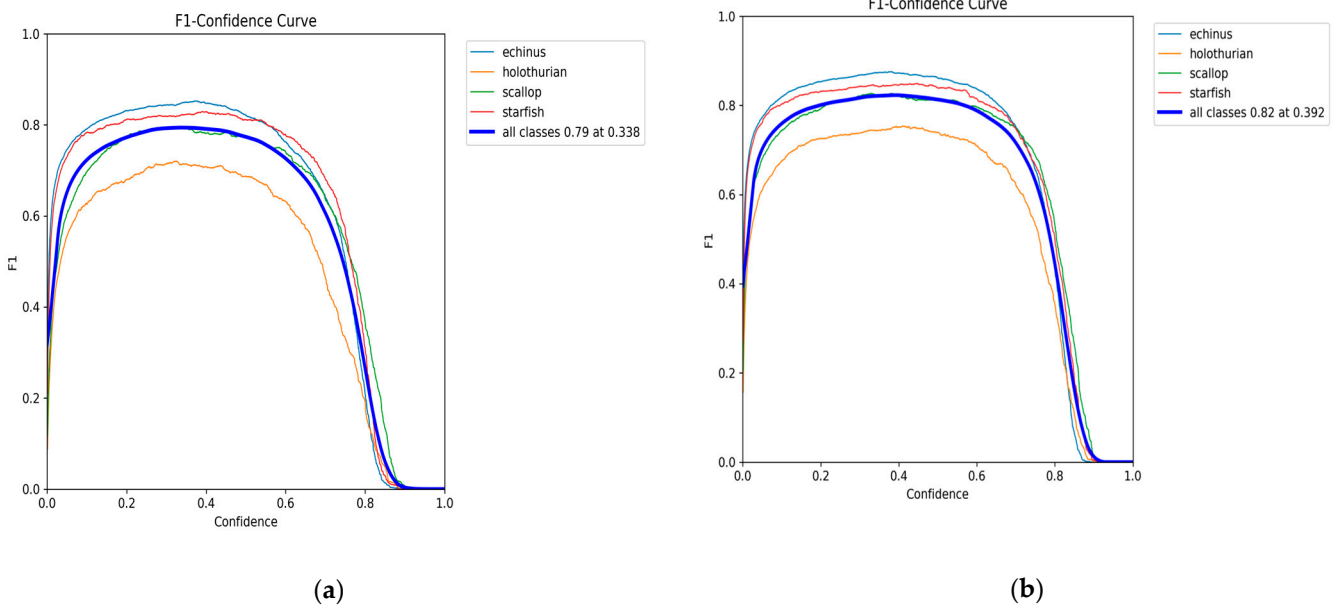


Figure 16. F1 curve. (a) The before effect; (b) the after effect.

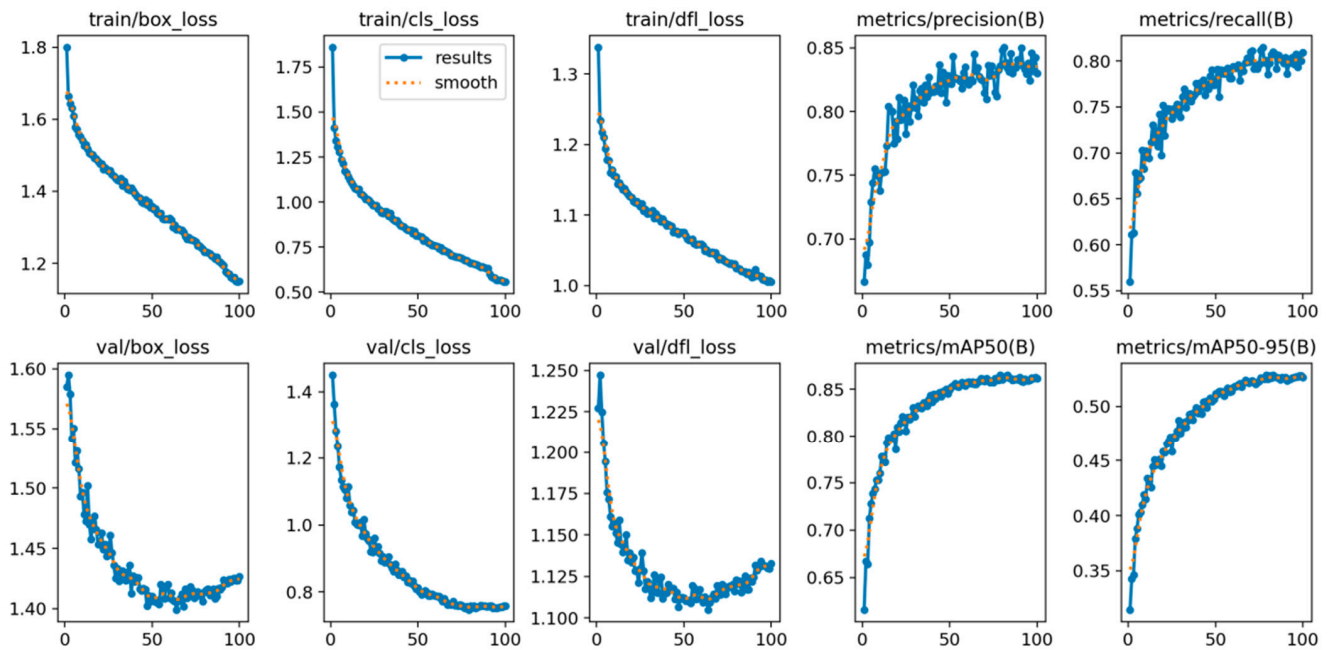


Figure 17. Parameter curve.

3.4.2. Evaluation of Several Object Detection Models' Detection Capabilities in Comparison

Previous research demonstrated that target identification techniques, including SSD, YOLOv5, RetinaNet, and Faster R-CNN, yield better detection results when used in underwater target detection. Thus, we used the same dataset to compare the YOLOv8s model with the Faster R-CNN, SSD, RetinaNet, and YOLOv5 models in order to assess the performance and advantages of the suggested techniques. The detection results are shown in Table 2.



Figure 18. Improved model validation set prediction graph.

Table 2. Results of the comparison of the AP, mAP, and FPS values with those of other main-stream models.

Model Name	AP (%)				MAP (%)	FPS (Hz)
	Sea Urchin	Sea Cucumber	Starfish	Scallop		
SSD	74.7	69.9	75.2	60.2	70.0	21
YOLOv5s	91.3	75.1	85.0	84.4	83.9	97
RetinaNet	77.2	68.1	78.3	61.2	71.2	26
Faster R-CNN	87.4	69.4	80.5	61.3	74.4	12
YOLOv8s	90.1	74.7	87.3	85.5	84.4	96
Improve YOLOv8s	92.1	76.5	90.2	87.2	86.5	85

The enhanced YOLOv8s method suggested in this paper for underwater object identification tasks outperforms the widely used two-stage object detection network, Faster RCNN, in terms of detection speed and accuracy, as indicated by the data in the table. Meanwhile, the benefits of the suggested model are further supported by the improved YOLOv8s algorithm’s notable advantages in mAP.

3.4.3. Ablation Experiment

The results of ablation tests are presented in this document to verify the functionality of many of the submodules. By progressively adding many modules and assessing each module’s enhanced impact on the overall performance of the model using the ablation tests, the efficacy of the procedure suggested in this study was confirmed. The results of the experiment are shown in Table 3. B represents the results obtained by using the dataset in the YOLOv8s model, S represents the use of SPPCSPC in YOLOv8s, D represents the use of deformable convolution DCNv3 in YOLOv8s, and W represents the use of WIoU.

Table 3. Results of ablation experiments on underwater datasets.

Model	B	S	D	W	MAP (%)	FPS	FLOPs (G)	Parameter Quantity (M)
1	✓				84.9	96.3	28.4	11.1
2		✓			85.2	90.2	31.6	14.5
3		✓	✓		86.3	87.9	33.2	17.6
4		✓	✓	✓	86.5	85.7	33.2	17.6

The findings show that while Model 1’s FPS decreased by 6, Model 2’s floating-point operations (FLOPs) and parameter count increased somewhat. Compared with Model 4, Model 3 only changed the loss function, and the floating-point operations (FLOPs) and parameter count did not show significant changes. In comparison to Model 1, Model 4 exhibited a 1.6% rise in mAP, while FPS satisfied the actual demand. It is evident that the suggested approach works.

3.4.4. Underwater Robot Prototype Grasping Experiment and Result Analysis

(1) Artificial water tank experiment

Figure 19 shows the outcomes of the detection tests conducted in this artificial pool experiment. The experimental findings show that in addition to a large number of species, as well as frequent obstruction and overlap in multi-target recognition, the ROV can detect many targets at once.

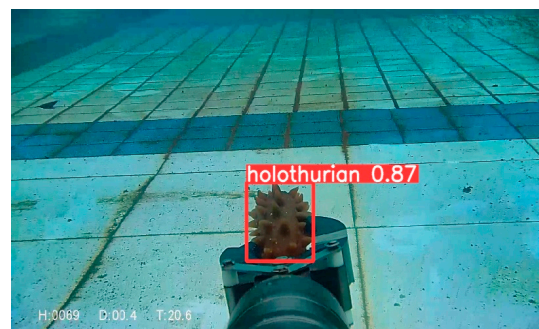


Figure 19. Target detection results of ROV pool.

(2) Natural water experiments

To validate the advantages of the study’s algorithm concerning detection effectiveness, typical scenes in various environments with complex underwater backgrounds, occluded environments, and multiple dense targets were selected. The network was tested and compared using YOLOv8s, as Figures 20–22 illustrate.

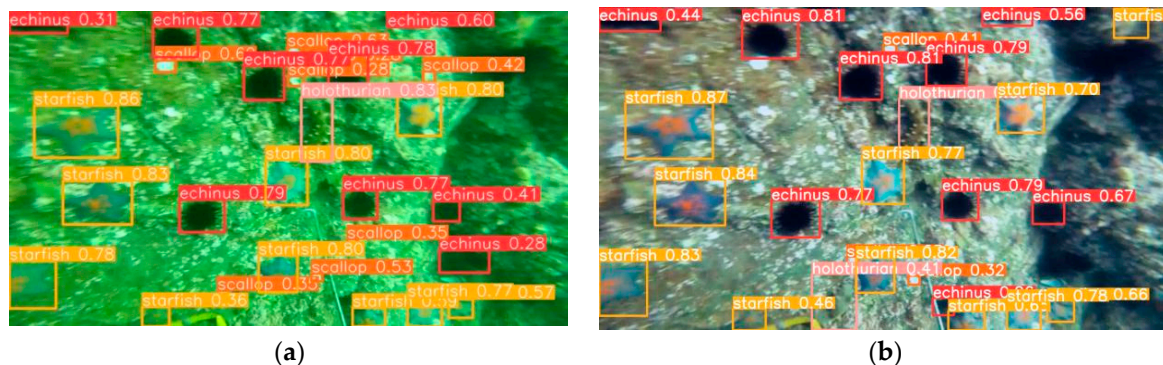


Figure 20. Comparison for complex background detection. (a) The before effect; (b) the after effect.

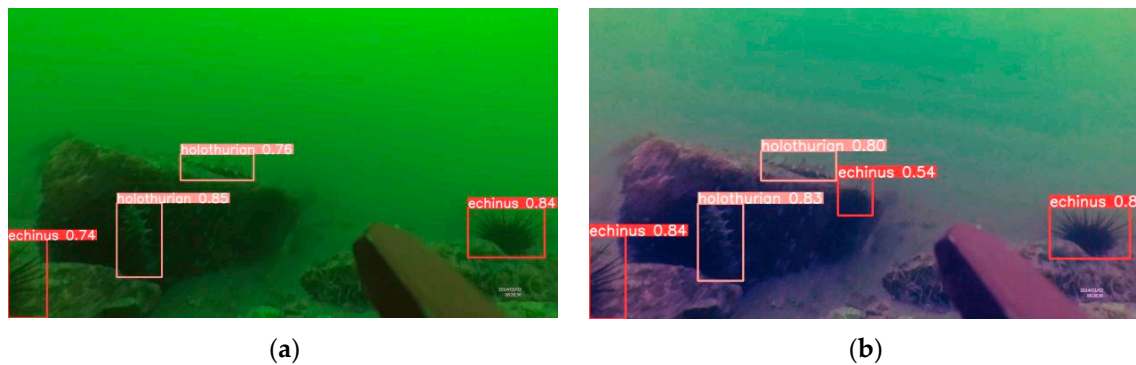


Figure 21. Comparison for occlusion environment detection. (a) The before effect; (b) the after effect.

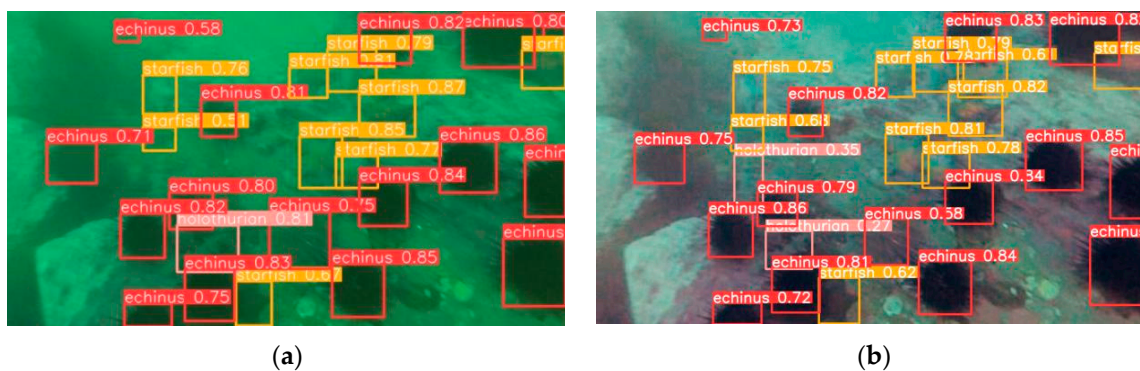


Figure 22. Comparison of underwater multi-class dense target detection. (a) The before effect; (b) the after effect.

The background of Figure 20 is mainly composed of rocks, which are similar in color to sea cucumbers. The white spots on the rocks resemble scallops. This background poses significant difficulties for the detection of sea cucumbers. The original YOLOv8s model network made detection errors in complex backgrounds, incorrectly detecting sea cucumbers with similar background colors. The improved YOLOv8s model can effectively solve this problem. The enhanced SPP structure may interact and fuse features from several phases in an efficient manner, increasing the model's detection accuracy. Figure 21 shows underwater photographs that are obstructed and fuzzy, and the initial model missed certain detections. However, the updated model is able to identify them with success. The enhanced model is more resilient and broadly applicable, since its deformable convolution can be closer to the size and shape of the item being sampled. When facing multiple dense images, as shown in Figure 22, the YOLOv8s network experiences missed detections, while the improved model can more accurately identify overlapping and small targets.

4. Conclusions

The present study presents an enhanced method for underwater robot object detection, utilizing YOLOv8s, in order to tackle the challenges of inadequate image quality and low recognition accuracy. Considering the issue of model parameters, only the convolution of the ninth layer is modified, and the deformable convolution is designed to be adaptive. Certain parts of the original convolution were replaced with DCN v3, in order to address the issue of the deformation of underwater photographs with fewer parameters and more effectively capture the deformation and intricate details of underwater targets. We utilized SPPFCSPC to improve multi-scale target recognition and to improve feature expression by combining high-level semantic features with low-level shallow features. Lastly, when WIoU loss v3 was substituted for the CIoU loss function, the overall performance of the model improved. According to the test results, the method suggested in this paper performs exceptionally well regarding underwater target recognition in challenging situations.

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

Funding: This research was funded in part by Postgraduate Research and Practice Innovation Program of Jiangsu Province (SJCX24_2489), and in part by the Jiangsu Province Industrial Prospect and Key Core Technology Project (BE2021135).

Data Availability Statement: The dataset used in the paper can be downloaded here: https://openi.pcl.ac.cn/OpenOrcinus_orca/URPC2020_dataset/datasets (accessed on 10 March 2022).

Acknowledgments: The authors thank the editors and anonymous reviewers for their critical comments and suggestions for improving the manuscript.

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

Abbreviations

ROV	Remotely Operated Vehicle
CNN	Convolutional Neural Network
R-CNN	Region-Based Convolutional Neural Network
Fast R-CNN	Fast Region-Based Convolutional Neural Network
Faster R-CNN	Faster Region-Based Convolutional Neural Network
SSD	Single Shot MultiBox Detector
YOLO	You Only Look Once
Resnet	Residual Network
SPPF	Spatial Pyramid Pooling-Fast
SPPFCSPC	Spatial Pyramid Pooling-Fast and Cross-Stage Partial Connection
WIoU loss	Weighted Intersection over Union Loss
CIoU loss	Complete Intersection over Union Loss
DCNv3	Deformable ConvNet v3
URPC	Underwater Robot Professional Challenge

References

- Chen, L.; Zheng, M.; Duan, S.; Luo, W.; Yao, L. Underwater target recognition based on improved YOLOv4 neural network. *Electronics* **2021**, *10*, 1634. [CrossRef]
- von Benzon, M.; Liniger, J.; Sørensen, F.F.; Pedersen, S. Investigation of operating range of marine growth removing rovs under offshore disturbances. *IFAC-PapersOnLine* **2022**, *55*, 85–90. [CrossRef]
- Mou, L.; Bruzzone, L.; Zhu, X.X. Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 924–935. [CrossRef]
- Khankeshizadeh, E.; Mohammadzadeh, A.; Moghimi, A.; Mohsenifar, A. FCD-R2U-net: Forest change detection in bi-temporal satellite images using the recurrent residual-based U-net. *Earth Sci. Inform.* **2022**, *15*, 2335–2347. [CrossRef]
- LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
- Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
- Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1137–1149. [CrossRef] [PubMed]
- 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; Part I 14; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37.
- 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.
- 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
- Wei, X.; Yu, L.; Tian, S.; Feng, P.; Ning, X. Underwater target detection with an attention mechanism and improved scale. *Multimed. Tools Appl.* **2021**, *80*, 33747–33761. [CrossRef]

12. 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](#)]
13. Li, Y.; Bai, X.; Xia, C. An improved YOLOV5 based on triplet attention and prediction head optimization for marine organism detection on underwater mobile platforms. *J. Mar. Sci. Eng.* **2022**, *10*, 1230. [[CrossRef](#)]
14. Lei, F.; Tang, F.; Li, S. Underwater target detection algorithm based on improved YOLOv5. *J. Mar. Sci. Eng.* **2022**, *10*, 310. [[CrossRef](#)]
15. Li, W.; Zhang, Z.; Jin, B.; Yu, W. A real-time fish target detection algorithm based on improved yolov5. *J. Mar. Sci. Eng.* **2023**, *11*, 572. [[CrossRef](#)]
16. Zhang, J.; Chen, H.; Yan, X.; Zhou, K.; Zhang, J.; Zhang, Y.; Jiang, H.; Shao, B. An improved yolov5 underwater detector based on an attention mechanism and multi-branch reparameterization module. *Electronics* **2023**, *12*, 2597. [[CrossRef](#)]
17. Li, K.; Wang, Y.; Hu, Z. Improved YOLOv7 for Small Object Detection Algorithm Based on Attention and Dynamic Convolution. *Appl. Sci.* **2023**, *13*, 9316. [[CrossRef](#)]
18. Liu, K.; Sun, Q.; Sun, D.; Peng, L.; Yang, M.; Wang, N. Underwater target detection based on improved YOLOv7. *J. Mar. Sci. Eng.* **2023**, *11*, 677. [[CrossRef](#)]
19. 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](#)]
20. Wang, C.-Y.; Liao, H.-Y.M.; Wu, Y.-H.; Chen, P.-Y.; Hsieh, J.-W.; Yeh, I.-H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 390–391.
21. Gao, H.; Zhu, X.; Lin, S.; Dai, J. Deformable kernels: Adapting effective receptive fields for object deformation. *arXiv* **2019**, arXiv:1910.02940.
22. Wang, W.; Dai, J.; Chen, Z.; Huang, Z.; Li, Z.; Zhu, X.; Hu, X.; Lu, T.; Lu, L.; Li, H. InternImage: Exploring large-scale vision foundation models with deformable convolutions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 14408–14419.
23. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable convnets v2: More deformable, better results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9308–9316.
24. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21002–21012.
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 2020 AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
26. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding box regression loss with dynamic focusing mechanism. *arXiv* **2023**, arXiv:2301.10051.

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.