Journal of
Marine Science and Engineering

MDPI

Article

Underwater Object Detection Algorithm Based on Adding Channel and Spatial Fusion Attention Mechanism

Xingyao Wang [1], Gang Xue [1], Shuting Huang [1,*] and Yanjun Liu [1,2,*]

[1] Institute of Marine Science and Technology, Shandong University, Qingdao 266237, China; 202120959@mail.sdu.edu.cn (X.W.); xuegangzb@163.com (G.X.)
[2] Key Laboratory of High Efficiency and Clean Mechanical Manufacture, Ministry of Education, Shandong University, Jinan 250061, China
[*] Correspondence: hst@sdu.edu.cn (S.H.); lyj111yjslw@163.com (Y.L.); Tel.: +86-152-7523-0512 (S.H.); +86-133-2513-6508 (Y.L.)

Abstract: Underwater target detection is the foundation and guarantee for the autonomous operation of underwater vehicles and is one of the key technologies in marine exploration. Due to the complex and special underwater environment, the detection effect is poor, and the detection precision is not high. In this paper, YOLOv5 (You Only Look Once v5) is used as the overall structural framework of the target detection algorithm, and improvement is made on the basis of its detection precision in the underwater environment. Specifically, an attention mechanism (Channel and Spatial Fusion Attention, CSFA) that fuses the channel attention and spatial attention is proposed and added to the YOLOv5 network framework, enabling the network to focus on both the prominent features of the detected object and the spatial information of the detected object. The proposed method was tested on the underwater target detection dataset provided by the China Underwater Robot Professional Competition. The experimental detection precision (P) reached 85%, the recall (R) reached 82.2%, and the mean average precision (mAP) reached 87.5%. The effectiveness of the proposed method was verified, and its underwater target detection performance was better than that of ordinary models.

Keywords: underwater target detection; YOLOv5; channel attention; spatial attention; attention mechanism

check for updates

Citation: Wang, X.; Xue, G.; Huang, S.; Liu, Y. Underwater Object Detection Algorithm Based on Adding Channel and Spatial Fusion Attention Mechanism. *J. Mar. Sci. Eng.* **2023**, *11*, 1116. https://doi.org/10.3390/jmse11061116

Academic Editor: Fabio Bruno

Received: 26 April 2023
Revised: 15 May 2023
Accepted: 21 May 2023
Published: 24 May 2023

1. Introduction

Marine resources have become an important support for economic development. Therefore, countries worldwide have focused their scientific research on undersea technology, and underwater vehicles have become the main means of undersea work. Underwater vehicles can be used in technical fields, such as observation and survey work, seabed sampling, construction and maintenance of seabed facilities, and the laying and maintenance of seabed pipelines [1,2]. Autonomous underwater vehicles and remotely operated vehicles equipped with intelligent underwater target detection systems [3–5] play an important role in developing and protecting marine resources. Underwater target detection technology is the foundation and guarantee for autonomous underwater vehicles work. However, the complexity of the underwater environment and problems such as image blurring caused by light absorption and scattering make the research on underwater target detection more challenging.

In recent years, many scholars have launched exploration and research on underwater target detection algorithms, part of scholars' focus on optical image processing. For example, Yang M et al. [6] systematically summarized a series of underwater image enhancement and restoration algorithms, Han M et al. [7] summarized intelligent defogging and color restoration algorithms for underwater images, and Liu R et al. [8] summarized a series of underwater image enhancement algorithms. Han J et al. [9] proposed a fast and efficient underwater image enhancement model based on conditional GAN with a good generalization ability using aggregation strategies and concatenate operations to take full advantage of the limited hierarchical features. Qi Q et al. [10] proposed an Underwater Image Co-enhancement Network (UICoE-Net) based on an encoder–decoder Siamese architecture. With the improvement of computers' GPU computing power, target detection systems based on neural networks have become the main research direction in computer vision [11], which can be divided into two-stage and single-stage target detection frameworks. The two-stage target detection algorithm is divided into two steps: first, region proposal (RP) is performed, and then, sample classification is performed through convolutional neural networks, such as R-CNN [12], Fast-RCNN [13], and Faster-RCNN [14].

Yuan Hongchun et al. [15] proposed a detection model specifically for fish by improving the network structure of Faster R-CNN. Through two times of transfer learning training networks, the detection precision has been improved. However, the region proposal network (RPN) is still used, which is ineffective for processing hard samples. Many anchor boxes are used during sampling, and most samples are invalid or low-quality. Therefore, Song Shaojian et al. [16] proposed an underwater biological target detection method based on Mask R-CNN. First, hard samples are enhanced, and then, the image is enhanced using the multiscale retinal enhancement algorithm. Finally, underwater target detection is achieved based on Mask R-CNN combined with transfer learning. However, this method's sampling anchor box ratio is fixed, and the detection effect for targets with too large or too small aspect ratio is poor. Chen Yingyi et al. [17] proposed a fish detection method based on convolution neural networks, which improved the fish recognition rate. However, this network uses many pooling layers to scale images, which is easy to filter small target information, resulting in small target miss detection. Cai Zhaowei et al. [18] proposed the target detection method of Cascade RCNN, which improves the network's ability to perceive location information by continuously adjusting the threshold value of the training hyperparameter *IoU* (Intersection over Union). However, this method uses a three-level detector, with each layer performing repetitive feature extraction and image scaling operations, resulting in a huge amount of computation, affecting the model's rapid convergence and detection speed. Zeng Lingcai et al. [19] proposed a method to add the adversarial occlusion network (AON) to the standard Faster R-CNN detection algorithm called the Faster R-CNN-AON network. The detection accuracy of this network is improved compared to the standard Faster R-CNN network. Liu Jia et al. [20] proposed an underwater object detection algorithm based on Faster R-CNN. First, the Swin Transformer is used as the backbone network of the algorithm. Second, the deep and shallow feature maps are superimposed and fused by adding the path aggregation network. Third, online hard example mining makes the training process more efficient. Fourth, the ROI pooling is improved to ROI align, eliminating the two quantization errors of ROI pooling and improving the detection performance.

Different from two-stage target detection algorithms, a single-stage target detection algorithm has a simple structure and is fast. It can directly identify the classification of objects by extracting features from the network, such as a single shot multibox detector (SSD) [21] and You Only Look Once (YOLO) [22] series of networks. Compared to other target detection networks, improvements based on YOLO series networks are applied more widely. Xu Jianhua et al. [23] proposed an improved underwater target detection method based on YOLOv3 network, optimizing the network structure through multi-level fusion, optimizing clustering candidate boxes and other methods, and improving the overall precision to 75.1%. Mao Guojun et al. [24] improved the YOLOv4 network model by constructing a module embedded at the end of the YOLOv4 network to discriminate shallow marine organisms and accurately identify obscured targets, improving the detection precision. However, the detection speed has decreased due to adding a module and a certain amount of parameters. Chen Lingyu et al. [25] improved the recognition accuracy and speed by replacing the upsampling module with the deconvolution module and incorporating depthwise separable convolution into the YOLOv4 network structure. Lei

Fei et al. [26] improved the accuracy of underwater target detection by replacing the basic backbone network of YOLOv5 with the Swin Transformer, improving the path aggregation network (PANet) method for multiscale feature fusion, and improving the confidence loss function based on different detection layers. Qiang Wei et al. [27] proposed an underwater target detection method based on improved SSD (Single Shot MultiBox Detector), which uses feature pyramid network to adapt to target multiscale variation to a certain extent and has a good fusion effect for large- and medium-sized target features. In contrast, small target features are easy to lose, resulting in a low detection rate for small targets.

The attention mechanism originates from the study of human vision. When humans observe an image, they do not observe every pixel of the entire image but instead focus on specific parts according to their needs. Moreover, humans will learn from previously observed images where their attention should be focused when observing images in the future. Therefore, many scholars have applied attention mechanisms to computer vision research, utilizing limited visual information reasonably, selecting prominent features in the visual region, and then focusing on it. Adding an attention mechanism to a target detection network is an important direction in the research of underwater target detection algorithms. Zhao Xiaofei et al. [28] proposed FRANet (Feature Refinement and Attention Mechanism Network), which combines an anchor box trimming module, a spatial attention module, and a target detection module to form a cascade attention mechanism to solve the problem of obscured and classification imbalance. However, stacking multiple modules will inevitably increase computational consumption, and pooling in spatial attention will lose some distinguishing features. Wei Xiangyu et al. [29] combined SENet [30] (Squeeze and Extortion Networks) with YOLOv3 (You Only Look Once v3) [31] to enhance the semantic information of deep features and fuse them with shallow features across layers to address the problem of feature loss caused by fuzzy underwater disturbance and occlusion. This method still does not solve the problem of pooling causing a loss of detail. Zou Ziyin et al. [32] believed that labelling obscured objects in blocks would cause the network to be unable to pay attention to important areas, so they concatenated CBAM (Convolutional Block Attention Module) [33] and SENet to enhance spatial and channel features. CBAM itself is a concatenation of channel attention and spatial attention, and concatenation of SENet again can lead to the problem of channel information redundancy.

The channel attention mechanism obtains the importance of each feature channel through learning and focusing on the relationships between channels in the feature map, but it cannot capture feature information in the spatial dimension. The spatial attention mechanism believes that the contribution of each region in the image to the task varies, and the regions related to the task require special attention. Adding channel attention or spatial attention to underwater target detection algorithms can improve the detection accuracy. However, it can lead to losing spatial or channel feature information. Simply connecting the two attention mechanisms in series or parallel can lead to a more complex network structure, requiring more computing resources and higher computational complexity. On the other hand, it can also result in the separation of channel and spatial information, making it impossible to interact. Therefore, this paper proposes an attention mechanism that fuses channel and spatial attention, which can obtain channel and spatial feature information and facilitate cross-latitude information exchange. To address the challenges of difficult detection, easily missed detection, and false detection of underwater targets, an attention mechanism that fuses channel attention and spatial attention is added to the YOLOv5 underwater target detection algorithm to achieve high-precision detection of complex underwater scenes.

The other sections of this paper are as follows. Section 2 introduces the method proposed in this paper and the improved YOLOv5 network structure. In Section 3, the dataset and evaluation indicators were introduced. Section 4 is the experimental part of this paper and analyzes the experimental results. Finally, Section 5 concludes this paper.

2. Overview of Improved Network Structure

2.1. Channel and Spatial Fusion Attention Principle

The role of the channel attention mechanism is to obtain the importance of each channel in the feature map and, then, use this importance to give a weight value to each feature, thus letting the neural network pay attention to certain feature channels, enhance the channels of the feature map that are useful for the current task, and suppress the feature channels that are not useful for the current task. The role of the spatial attention mechanism is to obtain the importance of the location information of the feature map and use this importance to give a weight value to the feature, thus letting the neural network select important spatial regions or directly predict the most relevant spatial locations.

Channel and Spatial Fusion Attention believes that channel attention and spatial attention should not be simply connected in series or parallel but should interact with information across dimensions. This paper uses the channel split module to divide the input feature channel equally into two parts: one part for channel attention calculation first. The dimension $c_1 \times h \times w$ (h, w, and c_1 represent height, width, and number of channels) of the input feature map needs to be reduced to $c_1 \times 1 \times 1$, which is achieved through global pooling. Then, the obtained dimension $c_1 \times 1 \times 1$ is integrated into the fully connected layer to learn the importance of each channel. Finally, after being activated by the sigmoid function, different weights are assigned to the channels of the input feature map through the scale operation. The other part for the spatial attention calculation first performs average pooling and maximum pooling separately from the channel dimension and adjusts the dimension from $c_2 \times h \times w$ reduced to $1 \times h \times w$, and then, they merge to obtain a convolutional layer with a channel number of 2 ($2 \times h \times w$). A spatial attention with a channel number of 1($1 \times h \times w$) is obtained through another convolution. Finally, after being activated by the sigmoid function, different weights are assigned to the spatial dimensions of the input feature map through the scale operation. Then, a concat module is used to connect the two parts after the calculation, and finally, a channel shuffle [34] module is used for information interaction. The structure of Channel and Spatial Fusion Attention is shown in Figure 1.

Currently, convolutional neural networks are composed of multiple blocks with the same structure. ResNeXt [35] and MobileNet [36] propose that depthwise separable convolution and group convolution achieve a trade-off between precision and computational cost. However, these networks are not fully used 1×1 convolution (referred to as pointwise convolution in MobileNet) because 1×1 convolution requires considerable complexity. In order to solve this problem, the most direct method is to add group convolution to the 1×1 convolution, which will significantly reduce the computational complexity of the convolution.

Of course, group convolution also has certain disadvantages because general convolutions always do full channel convolutions on input feature maps, a channel-dense connection method. However, group conversion is a channel-sparse connection method. Group convolution groups different feature maps of the input layer and, then, uses different convolution kernels to convolve each group. The feature maps between different groups do not communicate with each other, therefore reducing the network feature extraction ability. As shown in Figure 2a, there are two group convolutions, GConv1 and GConv2. The pink channel always processes only the pink information, and the yellow and green parts are the same. Obviously, the output of a group is only related to the input within the group, which causes the information flow between channel groups to be unable to flow and weakens the information representation.

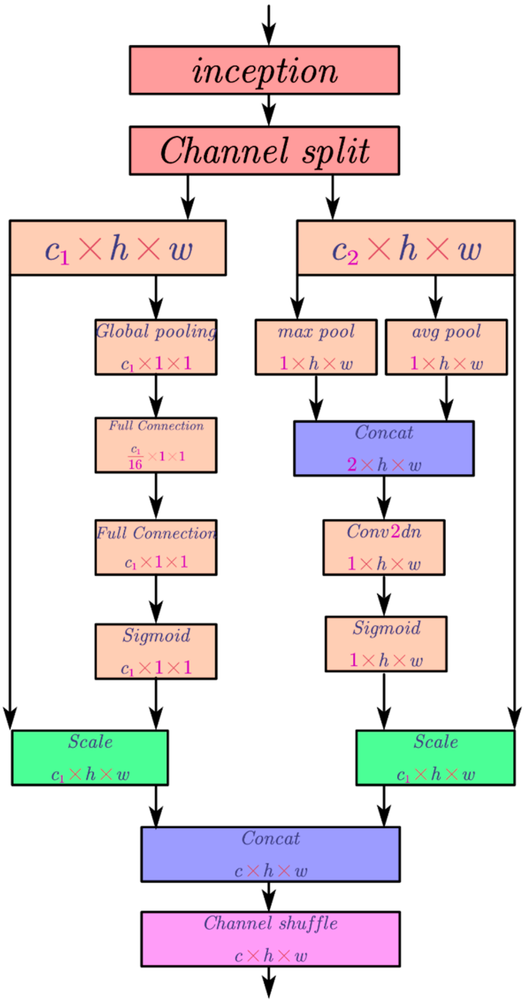


Figure 1. Structural diagram of Channel and Spatial Fusion Attention.

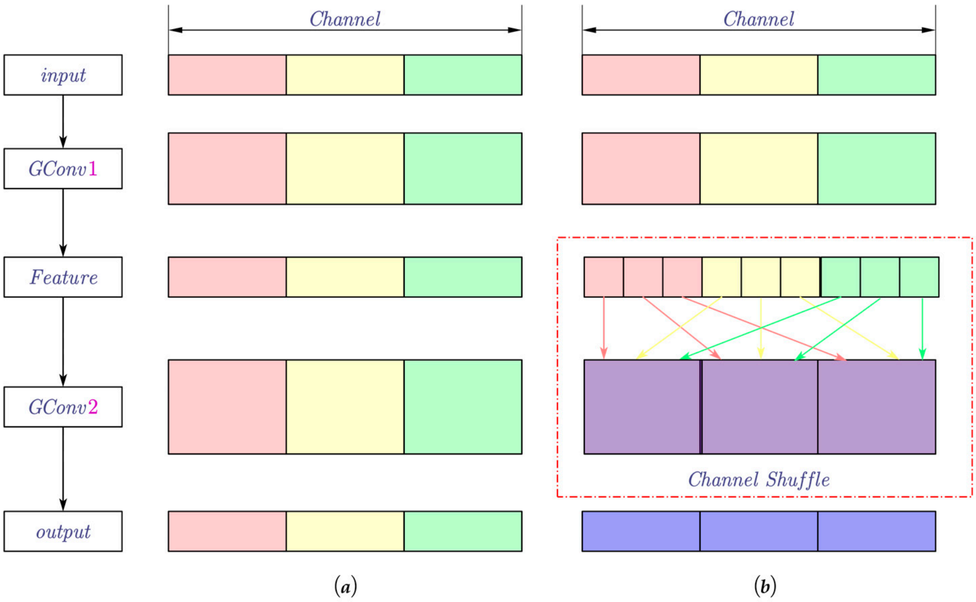


Figure 2. Structural diagram of Channel Shuffle: (**a**) Group convolution; (**b**) Group convolution with Channel Shuffle (One color represents a group convolution).

In order to solve the side effects caused by group convolution, the channel shuffle was proposed to help information flow between channels. As shown in Figure 2b, its meaning is to "reorganize" the feature map after the group convolution, which can ensure that the input of the group convolution to be used next comes from different groups so that information can flow between different groups. This process is not random but rather "evenly disrupted".

The channel shuffle operation is shown in Figure 3. Assume a convolution layer with a group of g whose output has g × n channels; first, reshape the output channel dimension to (g; n), then perform a transpose operation, and finally, flatten it as the input for the next layer. In addition, the channel shuffle is differentiable, which means it can be embedded in the network structure for end-to-end training.

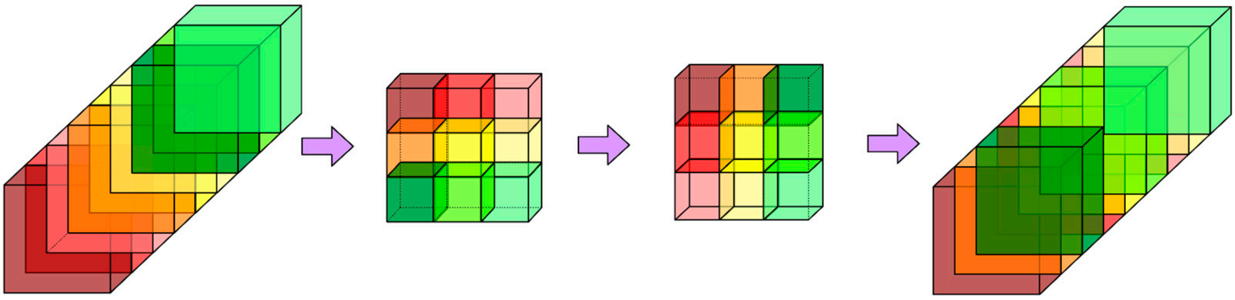


Figure 3. Channel shuffle operation diagram (One color represents a group convolution and different shades of the same color represent a channel).

2.2. Improved YOLOv5 Network Structure

Glenn Jocher released YOLOv5 [37] in 2020. This paper adds an attention mechanism based on the YOLOv5 network structure. The YOLOv5 network comprises multiple modules, which can be divided into the backbone network for extracting basic features of the target and the YOLO head for further enhancing features and making predictions. The backbone network of YOLOv5 is composed of the CSPDarknet53 network composed of multiple residual convolution blocks, which is part of the whole model with the largest amount of parameters, and its ability to extract features is related to the detection precision of the whole model. The improvement of YOLOv5 in this paper is to replace the Bottleneck module in the CSP structure with the attention module. The improved YOLOv5 network structure is shown in Figure 4.

YOLOv5 uses the SPP [38] (Spatial Pyramid Pooling) module, FPN [39], and PANet [40] module to extract features from three effective feature layers. SPP uses pooling at scales 13 × 13, 9 × 9, and 5 × 5 to increase the receptive field. Based on the feature pyramid, PANet performs repeated upsampling and downsampling of feature maps of different sizes, changing the original addition operation of PAN into connect, aiming to enhance the diversity of features further, improve the robustness of the model, and thereby improve the ability of the network to extract information. Finally, YOLOv5 uses three different scale anchor boxes to limit the range of prediction objects, thereby achieving the purpose of multiscale learning.

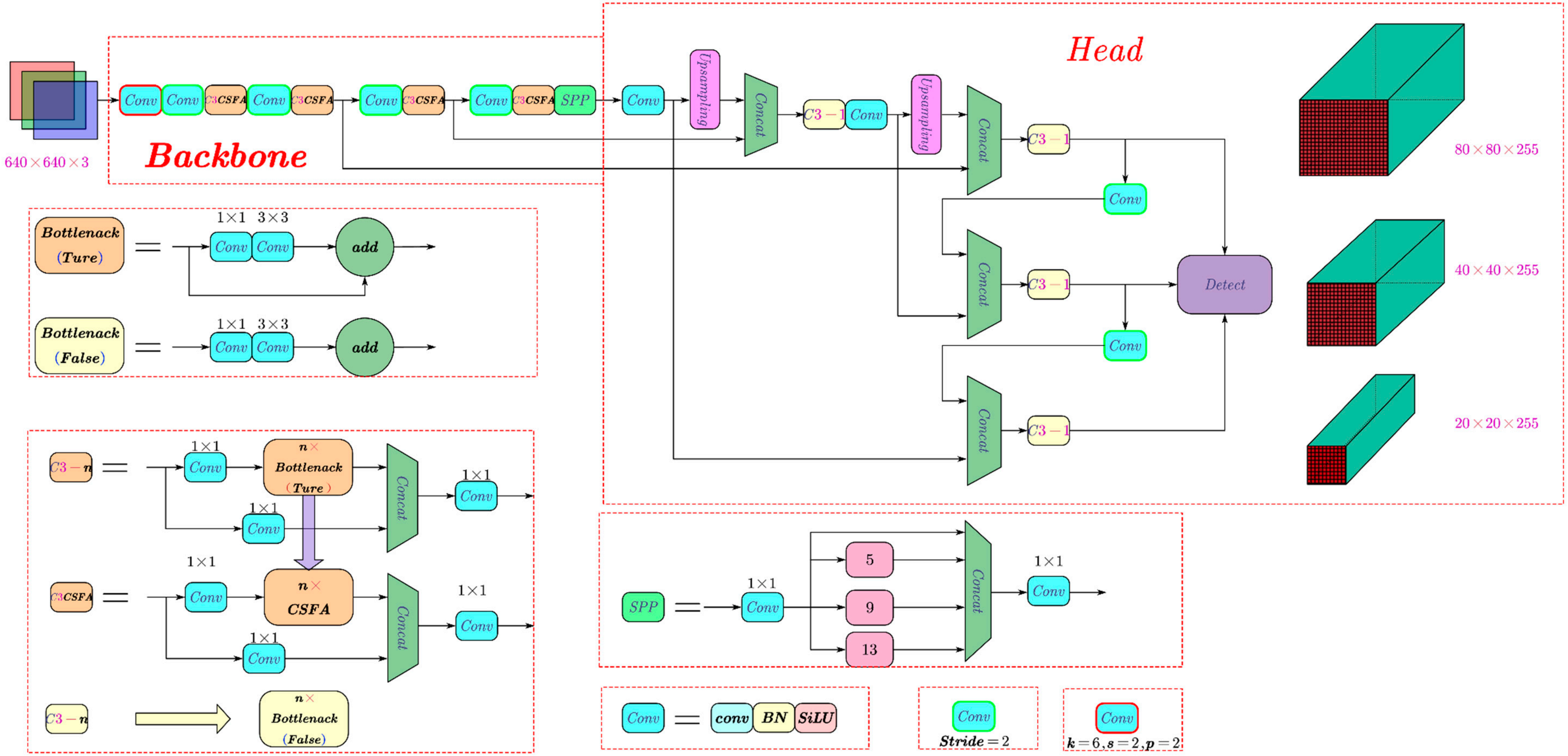


Figure 4. Improved YOLOv5 network structure.

3. Model Analysis

3.1. Data Set

The experimental dataset was from the Target Recognition Group of China Underwater Robot Professional Competition (URPC), with 6575 underwater target images in four classifications, as shown in Figure 5, including echinus, scallop, starfish, and holothurian. The dataset distribution is shown in Figure 6. Figure 6a shows the distribution of the number of detected targets. It can be seen from the figure that there are four classifications in total. The vertical coordinate represents the number of labelled images for a certain class. Therefore, 7899 echini, 1889 scallops, 2922 starfishes, and 1768 holothurians were labelled in the training set, totaling 14,478 target objects. All sample images were labelled using the Labelimg data label tool in the PASAL VOC sample set format, saved as the xml file, and then transformed to the yolo label format, which was saved as the txt file. Figure 6b shows the distribution of the size of the label box. As can be seen from the figure, the scale of the detected target was relatively wide, and the proportion of small targets was relatively large; Figure 6c shows the distribution of the center points of the normalized label box. The coordinate axes 0 to 1 represented the position of the normalized label box center coordinate points in the image. The center point coordinates in the figure covered the entire image from 0 to 1. It can be seen that the distribution range of the center coordinates (x, y) of the dataset were relatively wide, conforming to the characteristics of the random distribution of underwater target features in the image. Figure 6d shows the height and width distribution of the normalized label box. It can be seen that the distribution was concentrated on smaller values, and the distribution was most dense when the proportion reaches 0.0 to 0.2, indicating that the proportion of the target in the image was relatively small. In order to enable the designed model to learn the dataset fully, this experiment selected 70% of them as the training set and 30% as the test set. The training set contained 4538 images, and the test set contained 2037 images.

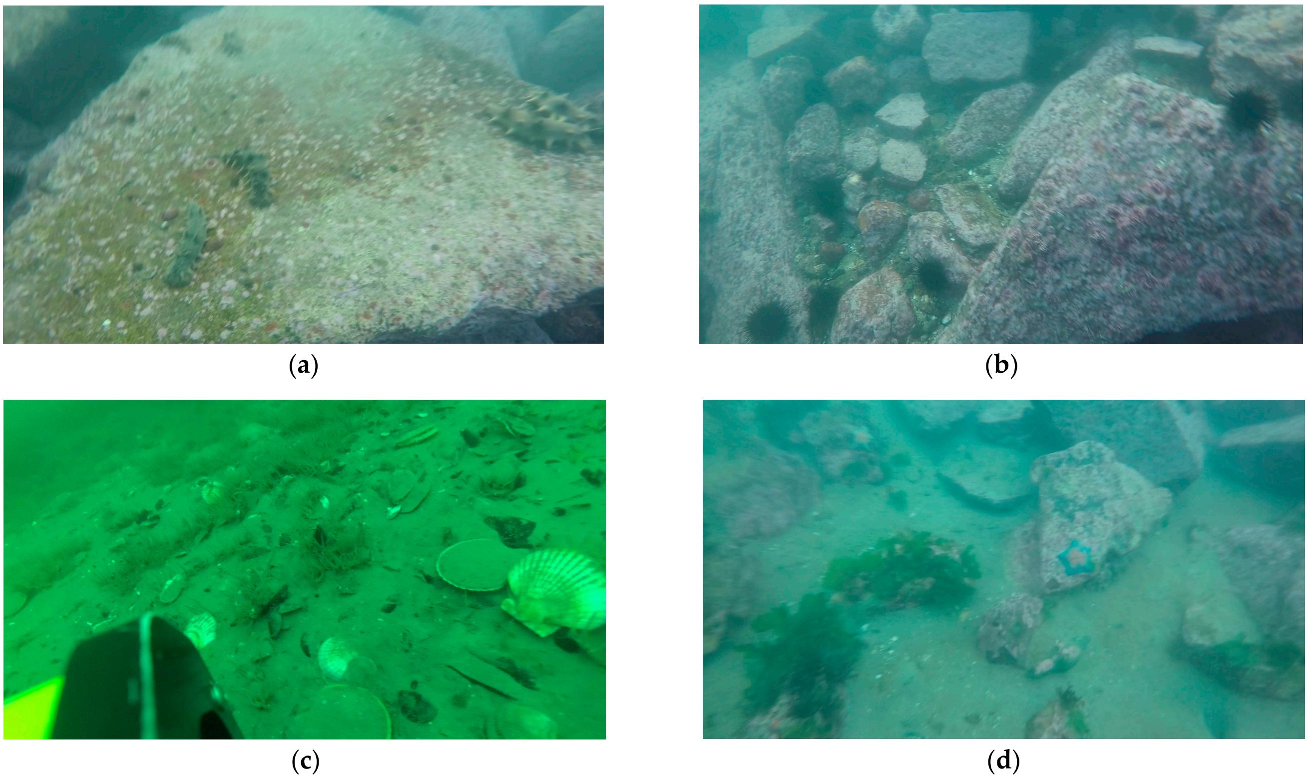
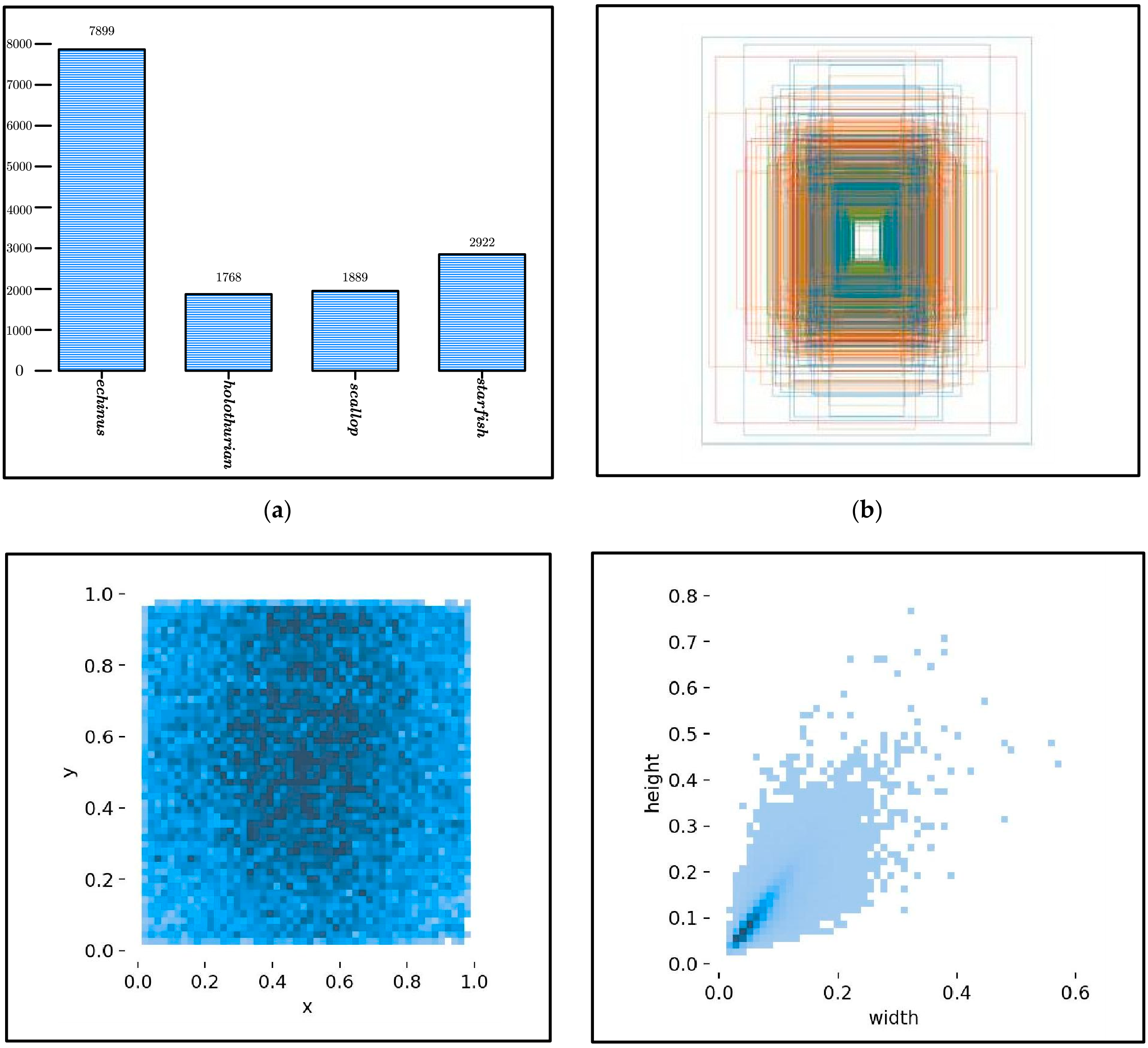


Figure 5. The dataset contains four classifications, which are (**a**) holothurian, (**b**) echinus, (**c**) scallop, and (**d**) starfish.



(**a**) (**b**)

(**c**) (**d**)

Figure 6. Dataset distribution: (**a**) Distribution of the number of targets of each classification; (**b**) Distribution of the label box size; (**c**) Distribution of the center points of the normalized label box; (**d**) The height and width distribution of the normalized label box.

3.2. Evaluation Metrics

There are two types of results for detecting underwater targets; one is to detect the correct target, and the other is to detect erroneous interfering objects, such as reefs. This paper used precision (P), recall (R), average precision (AP), and mean average precision (mAP) as evaluation metrics, as shown in Equation (1). IoU (Intersection over Union) was the ratio of the overlapping area and the combined area of the target predicted position (Box_r) and the practical target position (Box_t), which was used to measure the precision of target positioning. When IoU (experimental setting of 0.5) exceeded the set threshold value, the detection model considered Box_r a target location and marked it as TP. Otherwise, it was a non-target location, and the model marked it as FP.

$$IoU = (Box_t \cap Box_r)/(Box_t \cup Box_r), \tag{1}$$

The precision (P) formula is as follows:

$$P = \frac{TP}{TP + FP}, \quad (2)$$

in the formula, *TP* (True Positive) represents the number of samples that correctly detect underwater targets; *FP* (False Positive) represents the number of samples that incorrectly detect underwater targets. The recall (*R*) formula is as follows:

$$R = \frac{TP}{TP + FN}, \quad (3)$$

FN (False Negative) represents the number of samples where the target has not been detected. The mean average precision (*mAP*) formula is as follows:

$$mAP = \frac{\sum_{j=1}^{c} AP}{C}, \quad (4)$$

$$AP = \frac{\sum_{i=1}^{n} P}{N}, \quad (5)$$

The *AP* value specifically represents the average value of the correct probability of prediction for each class. *N* represents the total number of images containing target features, *P* represents the probability of correct prediction of target features in each image, and $\sum$ represents the sum of correct prediction probabilities for each target. *mAP* is the average of *AP* for all classes.

3.3. Model Training

Due to significant changes made to the YOLOv5 network model, it was necessary to validate the convergence of three main loss functions, including anchor box loss (Box), confidence loss (Objectness), and class loss (Classification). The situation after iterating 500 epochs is shown in Figure 7.

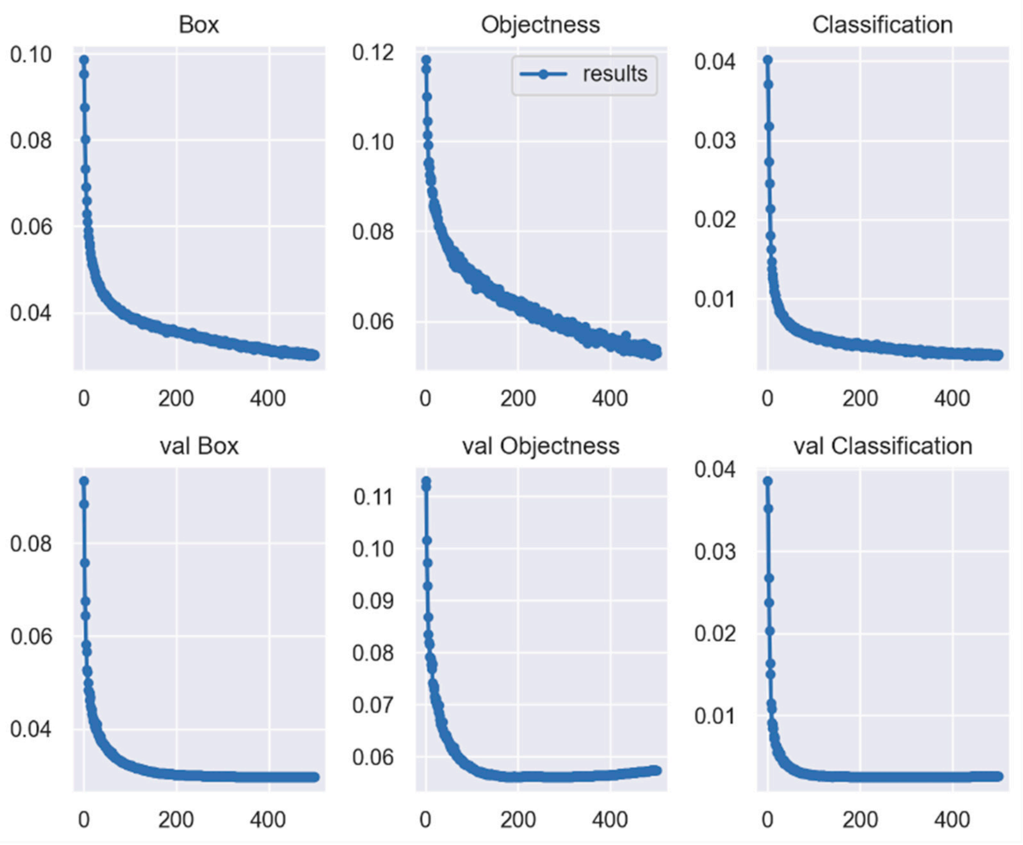


Figure 7. Convergence of the loss function.

From Figure 7, it can be seen that when training 500 epochs, Box, Objectness, and Classification could converge stably, with errors below 0.06, in the corresponding validation set (Val) loss function. The error of the three loss functions was also below 0.06, indicating that the model could stably converge.

4. Experimental Results

Experiments have verified the feasibility of improving the YOLOv5 target detection algorithm by adding an attention mechanism. The experimental results show that this method can improve the precision of target detection in complex underwater environments.

4.1. Experimental Environment

The underwater target detection dataset is divided into a training set and a test set in a 7:3 ratio. The basic parameter settings are shown in Table 1.

Table 1. Hyperparameter settings for network training.

Training Epochs	Batch Size	Learning Rate	Weight Decay	Momentum
500	16	0.01	0.005	0.9

The hardware environment of this experiment uses Intel (R) Core (TM) i9-11950H, CPU@2.60GH, and NVIDIA RTX A3000 with 32 G of memory. The programming uses Python 3.9. The model is optimized using the SGD (stochastic gradient descent) method. The network designed in this paper is trained and learned using a deep learning framework based on Pytorch. During the training process, the number of images per batch is 16, and the model is circulated through 500 training epochs in the dataset. The initial learning rate is set to 0.01, the weight decay is set to 0.0005, and the SGD momentum is set to 0.9. The mean average precision of underwater target detection is the evaluation metric for the model training results.

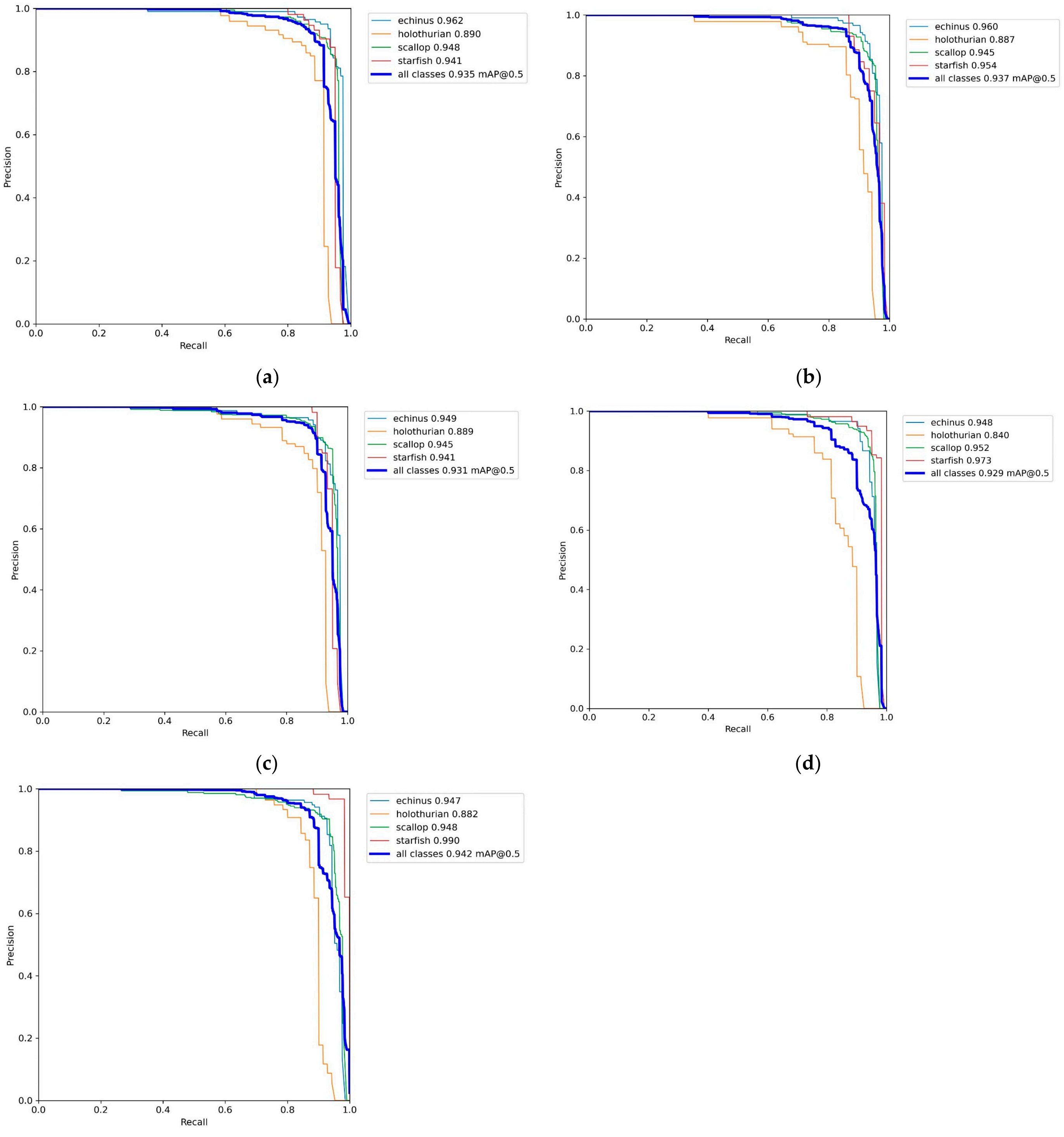
4.2. Ablation Experiment

In order to ensure the randomness of the experiment, 1110 images were randomly selected from the testset to test the YOLOv5 model and the improvement points of this paper. YOLOv5 represents that only the YOLOv5 model is used for experiments, YOLOv5+SE represents adding the SE attention mechanism to the YOLOv5 model, YOLOv5+CBAM represents adding the CBAM attention mechanism to the YOLOv5 model, and YOLOv5+CSFA represents adding the attention mechanism proposed in this paper to the YOLOv5 model, which fuses the channel attention and spatial attention. The test results of the above model are shown in Table 2 and the precision recall curve is shown in Figure 8.

As can be seen from Table 2, the precision of the YOLOv5 model is 91.2% while the precision of the model designed in this paper is 94.2%. The precision has improved, but the recall has decreased by 1.1 percentage points. Due to the addition of the model's channel attention mechanism and spatial attention mechanism and the increasing network layers, when learning similar targets to the bottom layer of the neural network, the feature differentiation is not large, resulting in the target being detected, but the class was incorrectly detected, leading to a reduced recall. The mean average precision is 0.7 percentage points higher than the YOLOv5, proving the improved model's effectiveness.

Table 2. Improvement point ablation experiment.

Model	P (%)	R (%)	mAP@0.5 (%)
YOLOv5	91.2	88.9	93.5
YOLOv5+SE	92	87.2	93.7
YOLOv5+CA	93	87.7	93.1
YOLOv5+CBAM	91.1	89.6	92.9
YOLOv5+CSFA(ours)	94.2	87.8	94.2



(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 8. Precision recall curve of ablation experiment: (**a**) YOLOv5, (**b**) YOLOv5+SE, (**c**) YOLOv5+CA, (**d**) YOLOv5+CBAM, and (**e**) YOLOv5+CSFA.

4.3. Comparison with Other Models

In order to objectively validate the effectiveness of the model, the YOLOv5+CSFA model and recent popular target detection models are compared on the dataset used in this paper. The results are shown in Table 3.

Table 3. Comparison results with other mod.

Model	AP (%, Holothurian)	AP (%, Echinus)	AP (%, starfish)	AP (%, Scallop)	mAP@0.5 (%)
RCNN [12]	68.2	80.4	78.9	69.3	74.2
Fast RCNN [13]	70.5	82.3	81.4	71.4	76.4
Faster RCNN [14]	74.1	85.5	84.4	75.2	79.8
YOLOv3 [31]	73.3	84.6	83.3	74.4	78.9
YOLOv3+SENet [29]	78	89.2	87.1	78.5	83.2
YOLOv4 [41]	78.2	90.7	86.4	78.3	83.4
YOLOv5 [37]	80	93.1	90.9	80.4	86.1
YOLOv6 [42]	78.4	93.5	91.5	78.8	85.5
YOLOv8 [43]	82.3	93.6	91.9	81.5	87.3
YOLOv5+CSFA(ours)	82.8	93.9	91.8	81.6	87.5

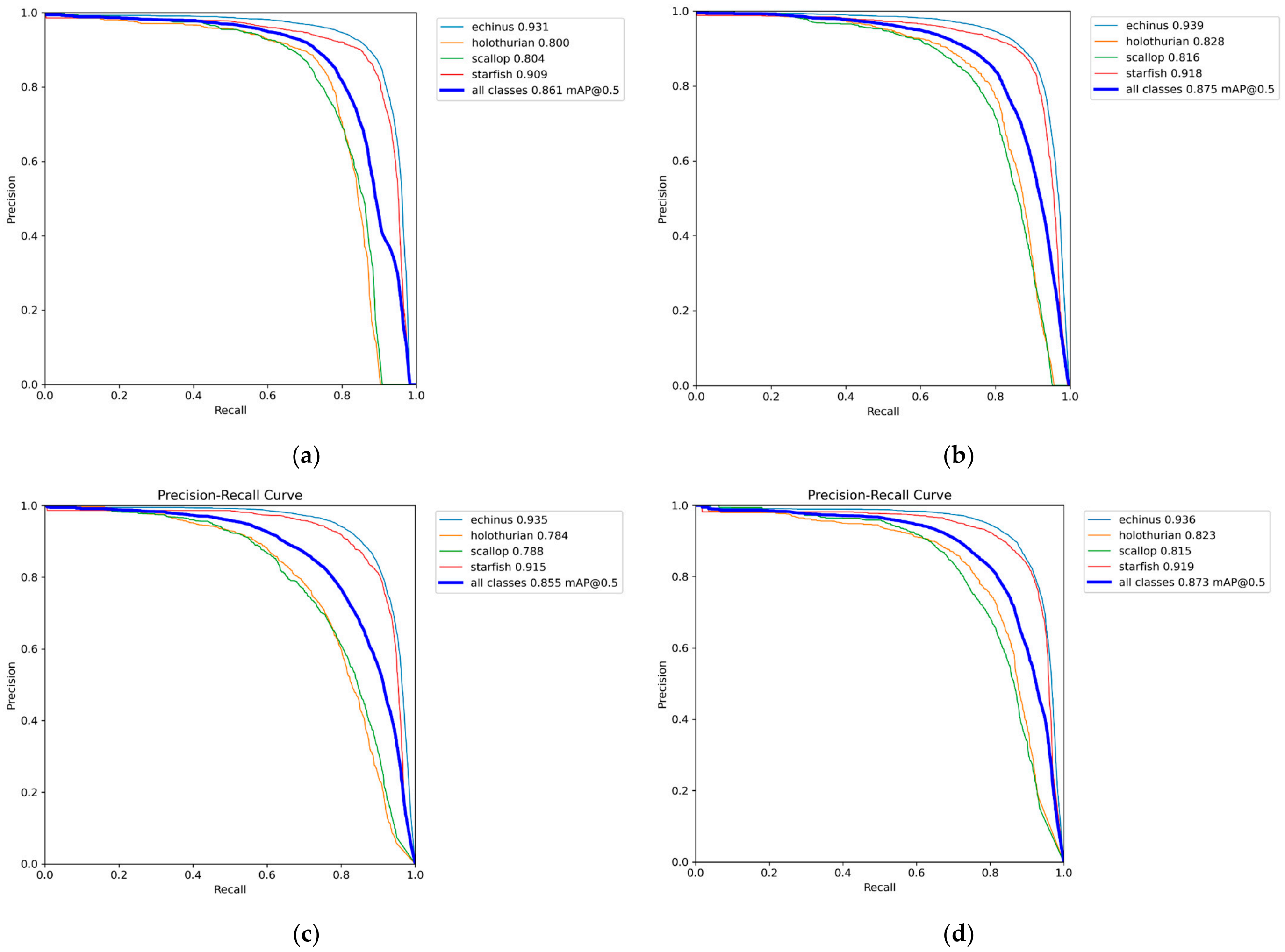
As can be seen from Table 3, compared to the popular two-stage target detection network, the mean average precision of the method proposed in this paper shows a significant advantage, which is 7.7 percentage points higher than the *mAP* of Faster RCNN. Compared with previous single-stage network YOLO (v3, v4) series models, the mean average precision is higher, which is 4.1 percentage points higher than the *mAP* of YOLOv4. Compared with the YOLOv5 underwater target detection network model in the same experimental environment and dataset, the mean average precision is 1.4 percentage points higher. Compared with the single-stage network YOLO (v6, v8) series models proposed in the past two years, the mean average precision is higher, which is 2 percentage points higher than the *mAP* of YOLOv6 and 0.2 percentage points higher than the *mAP* of YOLOv8, proving that the model has high precision in underwater target detection.

Figure 9 compares the precise recall curve between the improved YOLOv5 model and the YOLOv5 model, the YOLOv6 model, and the YOLOv8 model. As can be seen from the figure, the improved model achieves better detection results for all classes of targets. The *AP* value of the echinus reaches 90.5%, that of the starfish reaches 91.8%, that of the holothurian reaches 82.8%, that of the scallop reaches 81.6%, and the *mAP* value reaches 87.5%.

4.4. Grad-CAM Visualization

The cross-dimensional interaction provided by the channel and spatial fusion attention is assumed to facilitate the network learning for more meaningful internal representations of the images. The sample visualization of the Grad-CAM [44] technology was used to verify this statement, which visualizes the gradients of the top-class prediction concerning the input image as a colored overlay, as shown in Figure 10. Grad-CAM inputs the image into CNN, propagates it forward to obtain the first element (the last layer's output feature map), and obtains the model output's class logits (without softmax mapping). Then, it uses the class logit to be the certitude for backpropagation to obtain the gradient of the final layer of the output feature map concerning this class score. Finally, it calculates the average value of the spatial dimension of the feature map gradient to obtain the second element: a weight related to the class information and consistent with the number of channels in the feature map.

As shown in Figure 11, the channel and spatial fusion attention can capture tighter and more relevant bounds on the image of the underwater target detection dataset. In certain cases, when using the channel and spatial fusion attention, YOLOv5 can identify classes that the baseline model fails to predict correctly. These visualizations are beneficial for understanding the inherent ability of the channel and spatial fusion attention, which captures richer and more discriminative contextual information for specific target classes. This property of the channel and spatial fusion attention is extremely favourable and helpful in improving the performance of deep neural network architectures compared to their baseline counterparts.



(**a**) (**b**)

(**c**) (**d**)

Figure 9. Precision Recall Curve: (**a**) The YOLOv5 model; (**b**) Improved YOLOv5 model; (**c**) The YOLOv6 model; (**d**) The YOLOv8 model.

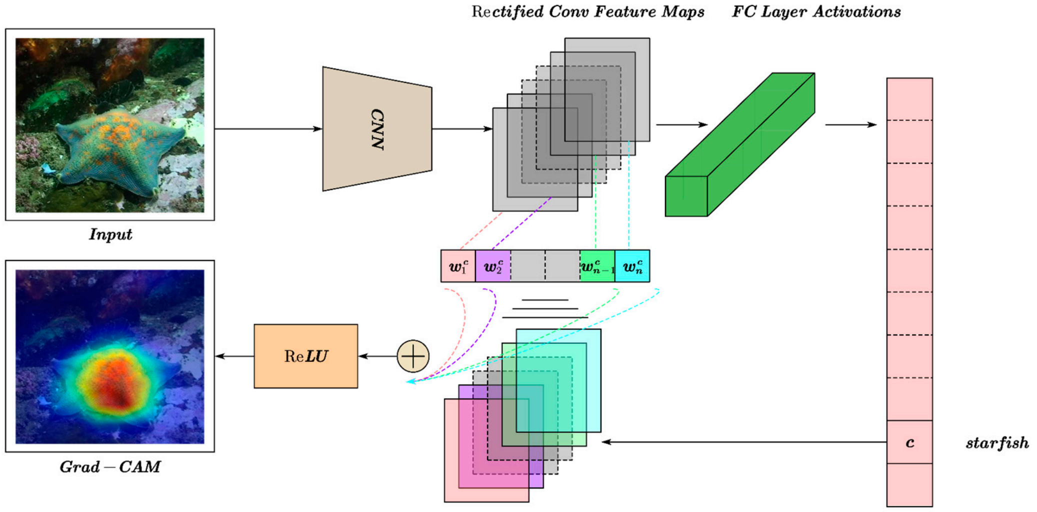


Figure 10. Grad-CAM Visualization.

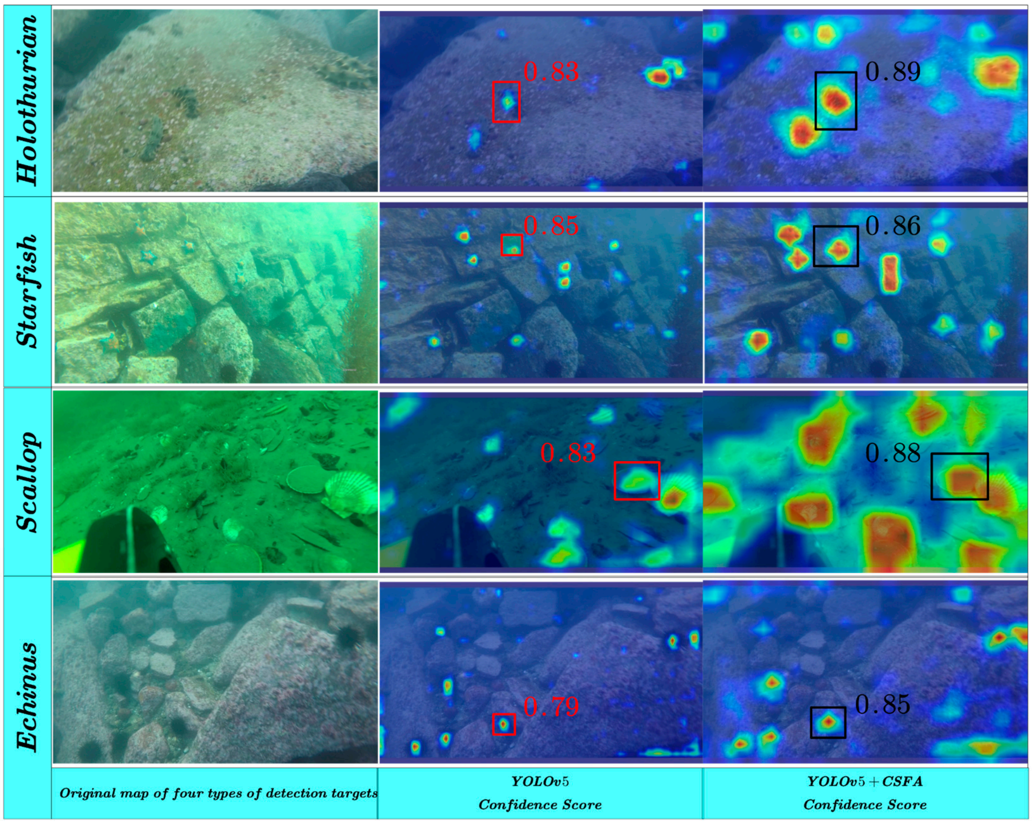


Figure 11. Obtain a random sample result for each class from the validation set and compare YOLOv5 and YOLOv5+CSFA.

4.5. Analysis of Detection Results

In order to more intuitively experience the detection results of the model, we randomly select 4 more complex images from the dataset and compare them with the YOLOv5 model. The selected original image is shown in Figure 12.

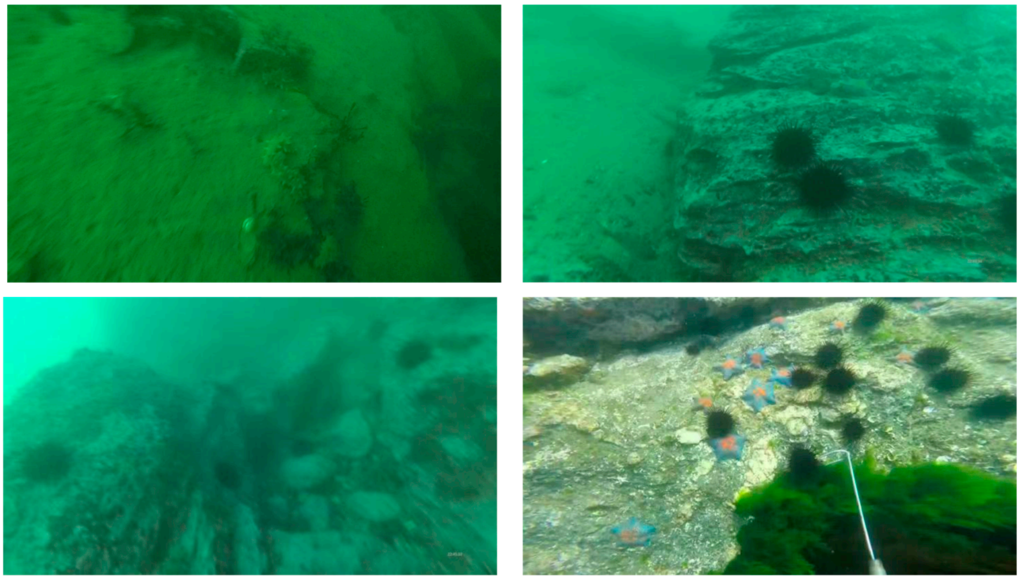


Figure 12. Original underwater target image.

As can be seen from Figure 12, echini, scallops, starfishes, and holothurians are relatively blurred in the water and highly fused with the background, making detection difficult. Figure 13 shows the detection results of the YOLOv5 model.

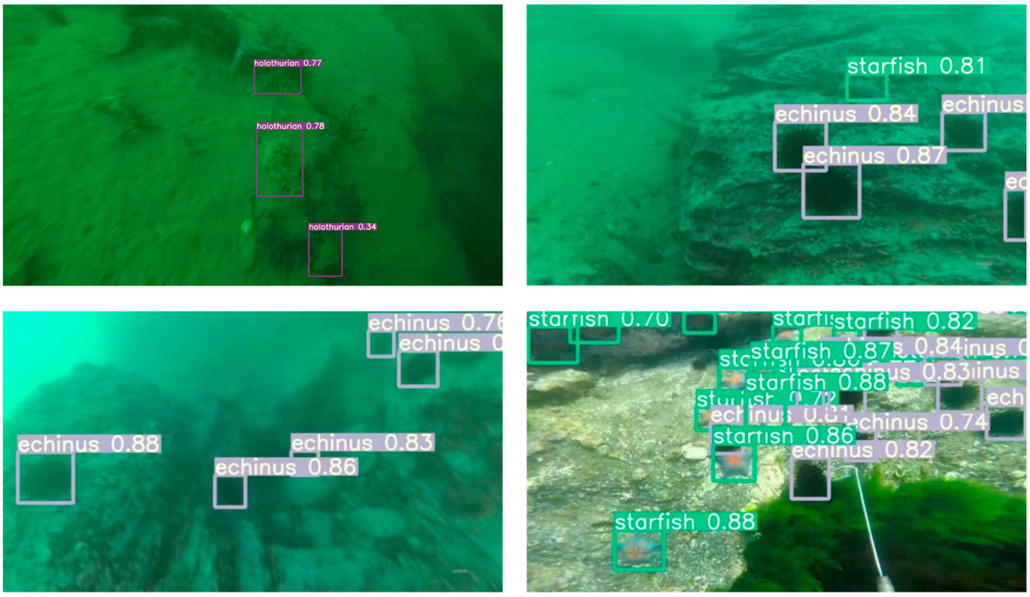


Figure 13. YOLOv5 Underwater Target Image Detection Results.

The YOLOv5 model correctly detects all target marine organisms. Figure 14 shows the detection results of the YOLOv5+CSFA model. By comparing the underwater target detection results of YOLOv5+CSFA designed in this paper with YOLOv5, it can be seen that the predicted values of YOLOv5+CSFA next to the prediction box are mostly higher than the YOLOv5 model, which proves the effectiveness of the improved model.

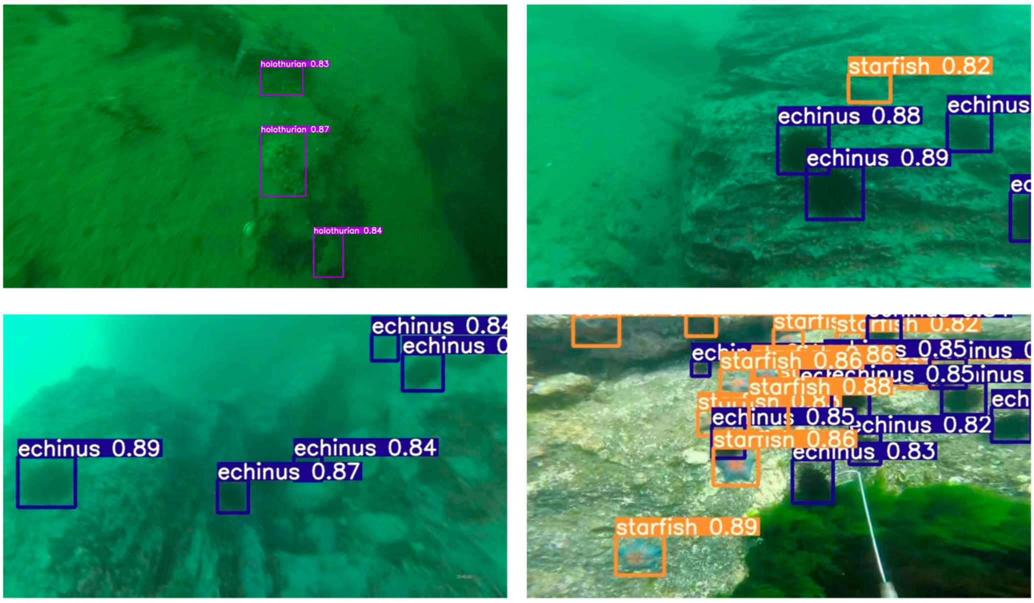


Figure 14. YOLOv5+CSFA Underwater Target Image Detection Results.

For the detection results of fuzzy and small targets, the comparison with YOLOv5 is shown in Figures 15 and 16:

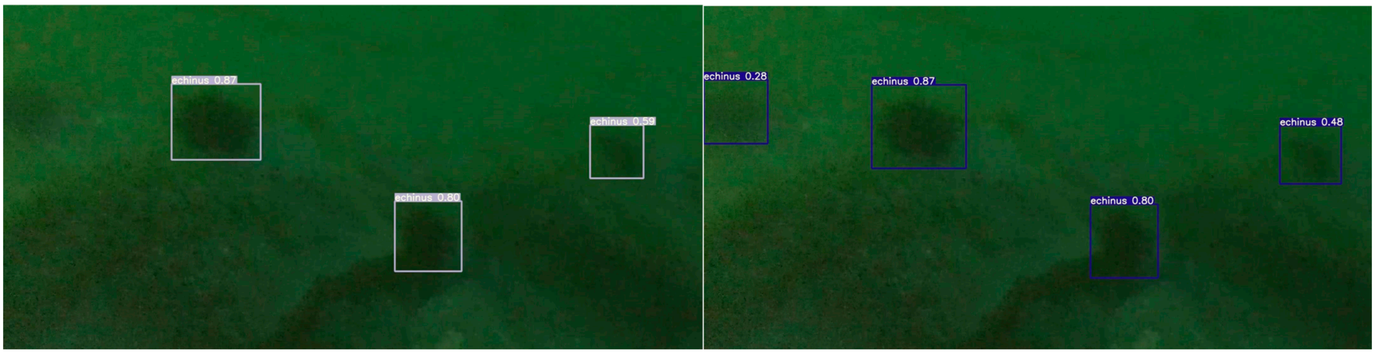


Figure 15. Comparison of Fuzzy Target Detection.

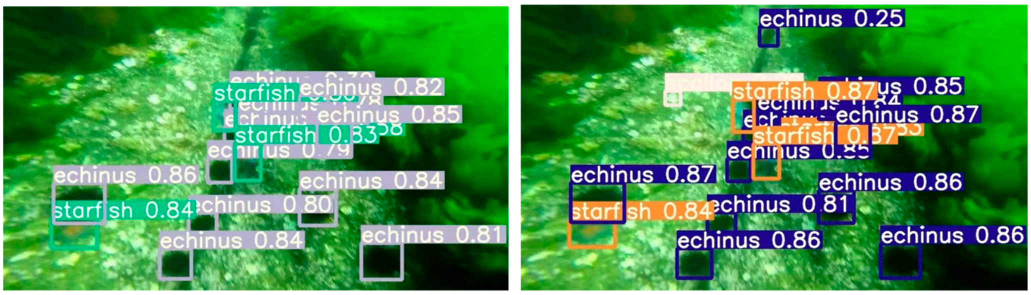


Figure 16. Comparison of Small Target Detection.

The left image in Figure 15 shows the detection results of YOLOV5, and the right image shows the detection results of YOLOV5+CSFA. It can be seen from the figure that YOLOv5+CSAF has detected the echinus that YOLOv5 failed to detect.

The left image in Figure 16 shows the detection results of YOLOv5, and the right image shows the detection results of YOLOv5+CSFA. The above image contains images of small underwater organisms. It can be seen that the right image has detected sea urchins that the left image failed to detect, and most predicted values are higher than those in the left. It has been proven that YOLOv5+CSFA is effective in detecting fuzzy and small underwater targets.

5. Conclusions

This paper improves the currently popular single-stage network YOLOv5 and applies it to the field of underwater target detection. In this paper, experiments were conducted on echinus, scallop, starfish, holothurian, and other difficult-to-capture undersea organisms, verifying the high precision of the detection model and expanding the application scenarios of underwater target detection algorithms. This paper compares the YOLOv5+CSFA model with popular target detection models in recent years, and the results show that the designed YOLOv5+CSFA model has a higher precision than other models. This paper conducted ablation experiments on the improved strategy. The experimental results showed that, compared to other attention mechanisms, the attention mechanism proposed in this paper that fuses the channel attention and spatial attention improves the target detection algorithm more significantly. The experimental results show that the detection results of the YOLOv5+CSFA model in complex underwater environments have been improved. The improved model is better than the general target detection model and is more robust in complex underwater environments.

Due to the complex underwater environment, data collection is relatively difficult, resulting in a slightly insufficient quantity and quality of the dataset. Therefore, part of our future work will focus on collecting datasets and improving their quality. The research on underwater image enhancement and restoration algorithms is one of the future research directions. Underwater target detection technology is the foundation and guarantee for achieving autonomous grasping operation of underwater manipulators. Therefore, another future research direction is to combine underwater target detection technology with deep reinforcement learning of underwater robotic arms to achieve autonomous grasping of underwater manipulators and conduct water tank experiments to verify grasping accuracy.

Author Contributions: Conceptualization, methodology, software, validation, data curation and writing—original draft preparation, X.W.; writing—review and editing, S.H. and G.X.; conceptualization, supervision and funding acquisition Y.L. All authors contributed to the design of the study. All authors have read and agreed to the published version of the manuscript.

Funding: The research was funded by the National Natural Science Foundation of China under the Project of Basic Science Center (42188102) "Monitoring/sampling/regulation technology and equipment for experimental water bodies of marine carbon sink process" and the Project of Shandong Province Key Research and Development Program (2019GHY112077) "Research on control characteristics of seven-axis redundant robotic arm system based on underwater mobile platform".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Thanks to Yanjun Liu for his guidance and support.

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

References

1. Liu, T.; Zhang, Q.; Zhang, Y.; Sun, Y.; Fan, Y. Dynamic Modeling and Simulation Analysis of Underwater Manipulator with Large Arms. *Chin. Hydraul. Pneum.* **2021**, *45*, 25–32.
2. Fan, Z.; Ha, Z. Underwater Manipulator Motion Spatial Analysis and Tracking Algorithm Optimization. *Mach. Electron.* **2020**, *38*, 67–73.
3. Sahoo, A.; Dwivedy, S.K.; Robi, P.S. Advancements in the field of autonomous underwater vehicle. *Ocean Eng.* **2019**, *181*, 145–160. [CrossRef]
4. Carlucho, I.; De Paula, M.; Wang, S.; Petillot, Y.; Acosta, G.G. Adaptive low-level control of autonomous underwater vehicles using deep reinforcement learning. *Robot. Auton. Syst.* **2018**, *107*, 71–86. [CrossRef]
5. Ullah, I.; Chen, J.; Su, X.; Esposito, C.; Choi, C. Localization and Detection of Targets in Underwater Wireless Sensor Using Distance and Angle Based Algorithms. *IEEE Access.* **2019**, *7*, 45693–45704. [CrossRef]
6. Yang, M.; Hu, J.; Li, C.; Rohde, G.; Du, Y.; Hu, K. An in-depth survey of underwater image enhancement and restoration. *IEEE Access* **2019**, *7*, 123638–123657. [CrossRef]
7. Han, M.; Lyu, Z.; Qiu, T.; Xu, M. A review on intelligence dehazing and color restoration for underwater images. *IEEE Trans. Syst. Man Cybern. Syst.* **2018**, *50*, 1820–1832. [CrossRef]
8. Liu, R.; Fan, X.; Zhu, M.; Hou, M.; Luo, Z. Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 4861–4875. [CrossRef]
9. Han, J.; Zhou, J.; Wang, L.; Wang, Y.; Ding, Z. FE-GAN: Fast and Efficient Underwater Image Enhancement Model Based on Conditional GAN. *Electronics* **2023**, *12*, 1227. [CrossRef]
10. Qi, Q.; Zhang, Y.; Tian, F.; Wu, Q.J.; Li, K.; Luan, X.; Song, D. Underwater Image Co-Enhancement with Correlation Feature Matching and Joint Learning. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 1133–1147. [CrossRef]
11. Gao, Y.; Yang, L.; Gao, P. Research on object detection algorithm based on yolov3. *Control. Instrum. Chem. Ind.* **2021**, *48*, 581–588.
12. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the 2014 the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
13. Girshick, R. Fast r-cnn. In Proceedings of the 2015 IEEE International Conference on Computer, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
14. 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**, *38*, 1137–1149. [CrossRef] [PubMed]
15. Yuan, H.; Zhang, S. An underwater fish object detection method based on Faster R-CNN and image enhancement. *J. Dalian Ocean. Univ.* **2020**, *35*, 612–619.
16. Song, S.; Zhu, J. Research on underwater biological object recognition based on mask R-CNN and transfer learning. *Comput. Appl. Res.* **2020**, *37*, 386–391.
17. Chen, Y.Y.; Gong, C.Y.; Liu, Y.Q. Fish recognition method based on FTVGG16 convolutional neural network. *Trans. Chin. Soc. Agric. Mach.* **2019**, *50*, 223–231.
18. Cai, Z.; Vasconcelos, N. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 1483–1498. [CrossRef]

19. Zeng, L.; Sun, B.; Zhu, D. Underwater target detection based on Faster R-CNN and adversarial occlusion network. *Eng. Appl. Artif. Intell.* **2021**, *100*, 104190. [CrossRef]
20. Liu, J.; Liu, S.; Xu, S.; Zhou, C. Two-Stage Underwater Object Detection Network Using Swin Transformer. *IEEE Access* **2022**, *10*, 117235–117247. [CrossRef]
21. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. *Eur. Conf. Comput. Vis.* **2016**, *14*, 21–37.
22. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
23. Xu, J.; Dou, Y.; Zheng, Y. An underwater object recognition and tracking method based on YOLO-V3 algorithm. *Chin. J. Inert. Technol.* **2020**, *28*, 129–133.
24. Mao, G.; Weng, W.; Zhu, J.; Zhang, Y.; Wu, F.; Mao, Y. Shallow sea biological detection model based on improved YOLO-V4 network. *Trans. Chin. Soc. Agric. Eng. (Trans. CSAE)* **2021**, *37*, 152–158.
25. Chen, L.; Zheng, M.; Duan, S.; Luo, W.; Yao, L. Underwater Target Recognition Based on Improved YOLOv4 Neural Network. *Electronics* **2021**, *10*, 1634. [CrossRef]
26. Lei, F.; Tang, F.; Li, S. Underwater Target Detection Algorithm Based on Improved YOLOv5. *J. Mar. Sci. Eng.* **2022**, *10*, 310. [CrossRef]
27. Qiang, W.; He, Y.; Guo, Y.; Li, B.; He, L. Research on underwater object detection algorithm based on improved SSD. *J. Northwestern Polytech. Univ.* **2020**, *38*, 747–754. [CrossRef]
28. Zhao, X.; Yu, S.; Li, Q.; Yan, Y.; Zhao, Y. Underwater object detection algorithm based on attention mechanism. *J. Yangzhou Univ. Nat. Sci.* **2021**, *24*, 62–67.
29. 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]
30. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef]
31. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1–12.
32. Zou, Z.; Gai, S.; Da, F.; Li, Y. Occluded pedestrian detection algorithm based on attention mechanism. *Acta Opt. Sin.* **2021**, *41*, 157–165.
33. 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; pp. 3–19.
34. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6848–6856.
35. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5987–5995.
36. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
37. *YOLO by Ultralytics*, version 5.0.0; Ultralytics: Washington, USA, 10 June 2020. Available online: https://github.com/ultralytics/yolov5 (accessed on 27 February 2023).
38. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *Eur. Conf. Comput. Vis.* **2014**, *37*, 346–361.
39. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
40. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
41. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
42. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. *arXiv* **2022**, arXiv:2209.02976.
43. *YOLO by Ultralytics*, version 8.0.0; Ultralytics: Washington, USA, 10 January 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 13 May 2023).
44. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.

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.