Article

Underwater-YCC: Underwater Target Detection Optimization Algorithm Based on YOLOv7

Xiao Chen 1,*, Mujiahui Yuan 1, Qi Yang 1, Haiyang Yao 1 and Haiyan Wang 1,2,*

1 School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi’an 710021, China; 211612111@sust.edu.cn (M.Y.)
2 School of Marine Science and Technology, Northwestern Polytechnical University, Xi’an 710072, China
* Correspondence: chenxiao@sust.edu.cn (X.C.); hywang@sust.edu.cn (H.W.)

Abstract: Underwater target detection using optical images is a challenging yet promising area that has witnessed significant progress. However, fuzzy distortions and irregular light absorption in the underwater environment often lead to image blur and color bias, particularly for small targets. Consequently, existing methods have yet to yield satisfactory results. To address this issue, we propose the Underwater-YCC optimization algorithm based on You Only Look Once (YOLO) v7 to enhance the accuracy of detecting small targets underwater. Our algorithm utilizes the Convolutional Block Attention Module (CBAM) to obtain fine-grained semantic information by selecting an optimal position through multiple experiments. Furthermore, we employ the Conv2Former as the Neck component of the network for underwater blurred images. Finally, we apply the Wise-IoU, which is effective in improving detection accuracy by assigning multiple weights between high- and low-quality images. Our experiments on the URPC2020 dataset demonstrate that the Underwater-YCC algorithm achieves a mean Average Precision (mAP) of up to 87.16% in complex underwater environments.

Keywords: underwater target detection; attention mechanism; YOLOv7; deep learning; model optimization

1. Introduction

The ocean is the largest repository of resources on Earth, and its related industries, such as marine ranching, are constantly improving due to the rapid development of underwater equipment. A crucial step in resource extraction and utilization is detection. New technologies, such as artificial intelligence, have provided significant impetus to improve detection. While many studies on underwater target detection are based on acoustic detection methods [1], these methods are inadequate for detecting small-sized underwater organisms due to their low sound source level, which can easily be drowned out by background noise. Additionally, the feature diversity in acoustic detection methods may not meet the demand for distinguishing small differences between underwater organisms. For this reason, optical images are more suitable for detecting small targets at close range, as they contain rich features of the target.

However, the complex underwater environment can seriously affect optical images. In general, the quality of underwater images is poor. The primary reason for this poor quality is the complexity and variability of underwater lighting conditions [2]. Specifically, (i) the energy attenuation of red to blue light in the chromatographic process changes from fast to slow, resulting in blue-green tone and underwater image color distortion. (ii) Different colors scatter in water to varying degrees and manners, causing loss of fine image details. (iii) Real-life water bodies are often turbid, containing sediment and plankton, which degrade the imaging quality of underwater cameras and blur the images. (iv) Due to the specific habitat of underwater organisms, they are usually attached to mud, sand, and reefs, which are difficult to distinguish from the background. Target occlusion is also a problem due to the specificity of organism distribution. All of these factors pose significant
challenges to underwater target detection, and traditional target detection algorithms are often less robust, more costly, and unsuitable for complex underwater environments [3].

Deep learning has demonstrated remarkable success in feature extraction, reducing the impact of errors caused by human factors. Its high speed and generalization make it widely used in many fields [4]. Deep-learning-based target detection algorithms can be broadly classified into two main categories. The first is a two-stage algorithm [5–8], which generates candidate regions on an image to determine if they contain a target. If a target is detected, the candidate region is classified with bounding box regression. However, the two-stage algorithm involves significant repetitive computation operations [9], leading to slow inference speed.

The one-stage detection algorithm is used to complete the target localization and regression directly on the image. OverFeat [10] was among the earliest one-stage detectors to be developed. Subsequently, the YOLO series [11–14] has demonstrated strong performance in practical engineering. In recent years, many researchers have applied YOLO networks to underwater target detection projects. Zhao et al. [15] proposed an underwater target detection algorithm, YOLO-UOD, based on YOLOv4-tiny. This algorithm introduced a symmetric FPN-attention module in the Neck architecture to achieve more efficient feature fusion and added a label-smoothing training strategy. This approach demonstrated superior detection performance. Zhang et al. [16] combined MobileNet V2 and depth-separable convolution to reduce the number of model parameters while using an improved AFFM for better fusion, achieving a balance between time and accuracy for underwater target detection. Li et al. [17] improved the feature extraction capability by embedding the triplet attention mechanism into the Neck structure of YOLOv5 and optimized the detection head to capture small-sized objects. This approach demonstrated good performance in detecting underwater organisms. Zhai et al. [18] added the CBAM module in YOLOv5s to save parameters and arithmetic power. They also increased the number of detection layers in the Head network by increasing the number of up-sampled layers in the Neck structure, thereby improving the accuracy of sea cucumber detection. Liu et al. [19] added CBAM to CSPDarknet53 to enhance the feature extraction of occluded and overlapping targets. Additionally, they used SAGHS to recover underwater images and finally obtain a detection model suitable for occluded underwater targets. Overall, these studies demonstrate the potential of YOLO-based algorithms for underwater object detection and the importance of optimizing network architectures and training strategies for specific applications.

In this paper, we propose a novel optimization algorithm, termed Underwater-YCC (YOLOv7 with CBAM and Conv2Former, YCC), for improving the accuracy of underwater target detection. Experimental results on the URPC2020 dataset demonstrate that Underwater-YCC outperforms YOLOv7 in terms of detection accuracy. The main innovations are as follows:

- Underwater data collection poses challenges due to the poor image quality and limited number of learnable samples. To overcome these challenges, this paper adopts data-enhancement methods, including random flipping, stretching, mosaic enhancement, and mixup, to enrich the learnable samples of the model. This approach improves the generalization ability of the model and helps to prevent overfitting.
- In order to extract more comprehensive semantic information and enhance the feature extraction capability of the model, we incorporate the CBAM attention mechanism into each component of the YOLOv7 architecture. Specifically, we introduce the CBAM attention mechanism into the Backbone, Neck, and Head structures, respectively, to identify the most effective location for the attention mechanism. Our experimental results reveal that embedding the CBAM attention mechanism into the Neck structure yields the best performance, as it allows the model to capture fine-grained semantic information and more effectively detect targets.
- To enhance the ability of the model to detect objects in underwater images with poor quality, this paper introduces Conv2Former as the Neck component of the network.
The Conv2Former model can effectively handle images with different resolutions and extract useful features for fusion, thereby improving the overall detection performance of the network on blurred underwater images.

- As low-quality underwater images can negatively affect the model's generalization ability, this paper introduces Wise-IoU as a bounding box regression loss function. This function improves the detection accuracy of the model by weighing the learning of samples of different qualities, resulting in more accurate localization and regression of targets in low-quality underwater images.

The paper is organized as follows. Section 2 focuses on the work related to this algorithm, with emphasis on the data enhancement approach and the YOLOv7 architecture. Section 3 introduces the content of the proposed Underwater-YCC algorithm. In Section 4 the relevant experimental results are analyzed and discussed. Section 5 presents conclusions.

2. Related Work

2.1. Underwater Dataset Acquisition and Analysis

Deep-learning models with good generalization ability require a substantial amount of training data, and a lack of appropriate data can lead to poor network training. The underwater environment is considerably more complex than the terrestrial environment, requiring the use of artificial light sources to capture underwater videos. Light transmission in water is subject to absorption, reflection, scattering, and other effects, resulting in significant attenuation. As a consequence, captured underwater images have limited visibility, blurriness, low contrast, non-uniform illumination, and noise.

The URPC2020 dataset is composed of 5543 images belonging to four categories: echinus, holothurian, scallop, and starfish. To train and test the proposed algorithm, the dataset was split into training and testing sets with an 8:2 ratio, resulting in 4434 images for training and 1109 images for testing. This dataset presents a variety of complex situations, such as underwater creatures gathering obscuration, uneven illumination, and motion-shot blurring, which makes it a realistic representation of the underwater environment and therefore will improve the generalization ability of the model. However, the uneven distribution of samples among categories and their different resolutions pose significant challenges to the model's training. Figure 1 shows the sample information of URPC2020.

Figure 1a shows the amount of data for each category, the size and number of bounding boxes, the location of the sample centroids, and the aspect ratio of the target occupying the entire image, respectively.

(a) (b)

Figure 1. The sample information of URPC2020; (a) labels, (b) example images.
2.2. Data Augmentation

Deep convolutional neural networks have demonstrated remarkable results in target detection tasks. However, these networks heavily rely on a large amount of image data for effective training, which is difficult to obtain in some domains, including underwater target detection. A detection model with high generalization ability can accurately detect and classify targets from various angles and in different states. Generalization ability can be defined as the difference in the performance of a model when evaluated on training and test data [20]. Models with weak generalization ability are prone to overfitting, and data augmentation is one of the key strategies to mitigate this issue and improve the generalization ability of the model.

2.2.1. Geometric Transformation

Geometric transformation is the alteration of an image and its inverse, such as flipping, rotating, shifting, scaling, cropping, etc. For orientation-insensitive tasks, flipping is one of the safest operations and the most commonly used, and it does not change the size of the target. In the case of underwater target detection, the movements, morphology, and orientation of underwater creatures are uncertain, and using the flip operation for data augmentation can effectively improve the training results of the model. Horizontal and vertical flips are the two most commonly used types of flip operations, and the horizontal flip is preferred in most cases.

2.2.2. Mixup Data Augmentation

The method of mixup data augmentation simply selects two random photos from each batch and mixes them in a certain ratio to generate a new image that is used in the training process, without the original image participating in the model training. It is a simple and data-independent data enhancement method that generates new sample-label data by adding two sample-label data images proportionally to construct virtual training examples [21]. The equation for processing data labels is as follows:

\[ \tilde{x} = \lambda x_i + (1 - \lambda)x_j \]  
\[ \tilde{y} = \lambda y_i + (1 - \lambda)y_j \]

where \( x_i \) and \( x_j \) are one-hot label encodings, \( y_i \) and \( y_j \) are one-hot label encodings; and \( (x_i, y_i) \) and \( (x_j, y_j) \) are two randomly selected samples in the training set, \( \lambda \in [0, 1] \). According to the above equation, mixup uses prior knowledge to extend the training distribution. Figure 2 shows the resulting graph after performing mixup data enhancement.

![Figure 2. Image enhancement with Mixup.](image)

2.2.3. Mosaic Data Augmentation

Mosaic data augmentation is a method of mixing and cutting four randomly selected images in a dataset to obtain a new image. The result contains richer target information, which
expands the training data to a certain extent and allows the network to be trained more fully on a small number of datasets. Figure 3 shows the image after mosaic enhancement.

Figure 3. Use of mosaic enhancement images during training.

2.3. Attention Mechanism

The attention mechanism can be regarded as a process of dynamic weight adjustment based on the features of the input image around the target position [22], so that the machine focuses on the target to be detected and recognized as much as possible, and optimizes the allocation of computing resources under limited computing power. Attentional mechanisms play an important role in the field of computer vision, and more and more people are optimizing models by introducing attentional mechanisms.

Attention mechanisms commonly used in the visual domain include the spatial domain, the channel domain, and the hybrid domain. The spatial domain is used to generate a spatial mask of the same size as the feature map. It then modifies the weights according to the importance of each location. The channel domain adds weight to the information on each channel, representing the relevance of that channel to the key information. The higher the weight, the higher the relevance. Finally, the hybrid domain effectively combines channel attention and spatial attention, allowing the machine to focus on both simultaneously. The attentional mechanisms can significantly improve the performance of target detection models.

Convolutional Block Attention Module

CBAM is a simple and effective feed-forward convolutional neural network attention module [23]. The CBAM combines a channel attention module with a spatial attention module, which has superior performance compared to attention mechanisms that focus on only one direction. Its structure diagram is shown in Figure 4. The features are first passed through a channel attention module, the output is weighted with the input features to obtain a weighted result, and then a spatial attention module is used for final weighting to obtain the output.
was proposed in 2022. It is faster and more accurate than the previous YOLO series and other methods. For the characteristics of underwater targets, we propose an optimization algorithm based on YOLOv7 to improve the detection accuracy of underwater organisms. The network structure of YOLOv7 is shown in Figure 7.

2.4. YOLOv7 Network Architecture

The YOLOv7 model [24] is a state-of-the-art, real-time, target-detection model that was proposed in 2022. It is faster and more accurate than the previous YOLO series and other methods. For the characteristics of underwater targets, we propose an optimization algorithm based on YOLOv7 to improve the detection accuracy of underwater organisms. The network structure of YOLOv7 is shown in Figure 7.

Figure 4. Structure of the CBAM.

The structure of the channel attention is shown in Figure 5. The input feature maps are subjected to w-based global max pooling and h-based global average pooling, respectively. The output is obtained after a shared fully connected layer is subjected to summation and Sigmoid activation operations to obtain the channel attention feature maps. \( M_c(F) \) represents the output feature maps of the channel attention mechanism.

\[
M_c(F) = \sigma(MLP(AvgPool(F) + MLP(MaxPool(F))))
\]  

Figure 5. Structure of the channel attention.

The spatial attention mechanism takes the output of the channel attention module as its input, performs channel-based global max pooling and global average pooling, concatenates the two results, reduces the dimensionality to a channel by a convolution operation, and then generates a spatial attention feature by sigmoid. \( M_s(F) \) represents the output feature map of the spatial attention mechanism. The structure of the spatial attention is shown in Figure 6.

\[
M_s(F) = \sigma(f^{7 \times 7}([AvgPool(F); MaxPool(F)]))
\]  

Figure 6. Structure of the spatial attention.
2.4. YOLOv7 Network Structure

The YOLOv7 network structure is a one-stage structure consisting of four parts: the Input Terminal, Backbone, Neck, and Head. The target image is fed into the Backbone after a series of operations for data enhancement. The Backbone section performs feature extraction on the image, the extracted features are fused in the Neck module and processed to obtain three sizes of features, and the final fused features are passed through the detection Head to obtain the output results. The Input Terminal involves features such as data enhancement, adaptive anchor box calculation, and adaptive image scaling; here, we will focus on the Backbone, Neck, and Head.

2.4.1. Backbone

The Backbone of the model is built using Conv1, Conv2, the ELAN module, and the D-MP module. Conv1 and Conv2 are two modules with different sizes of convolutional kernels, and the structure is shown in Figure 8, which is a convolutional layer superimposed with a batch normalization layer and an activation layer. Conv1 is mainly used for feature extraction, while Conv2 is equivalent to a down-sampling operation to select the features to be extracted.

ELAN is an efficient network structure that allows the network to learn more features by controlling the longest and shortest gradient paths and thus has better generalization capabilities. It has two branches. The first branch goes through a $1 \times 1$ convolution module...
to change the number of channels, the other branch changes the number of channels and then goes through four 3 × 3 convolution modules for feature extraction, and finally introduces the idea of residual structure to superimpose the features and attain more detailed feature information. The structure is shown in Figure 9.

Figure 9. The architecture of ELAN.

The D-MP module divides the input into two parts. The first branch is spatially down-sampled by MaxPool and then the channels are compressed by a 1 × 1 convolution module. The other branch compresses the channels first and then performs a sampling operation using Conv2. Finally, the results of both samples are superimposed. The module has the same number of input and output channels with twice the spatial resolution reduction. The structure is shown in Figure 10.

Figure 10. The architecture of D-MP.

2.4.2. Neck

The images go through the Backbone for feature extraction and then enter the Neck for feature fusion. The fusion part of YOLOv7 is similar to YOLOv5, using the traditional PAFPN structure. Three effective feature layers are obtained in the Backbone part for fusion. The features are first fused through an up-sampling operation and then through a down-sampling operation, thus obtaining feature information at different scales and allowing the network to have better robustness.

The SPPCSPC module first divides the features into two parts: one for conventional processing and the other for SPP operation. The features in SPP are passed through four different MaxPool modules with pooling kernels of 1, 5, 9, and 13, respectively; the maximum pooling is used to obtain different perceptual fields that are used to distinguish between large and small targets. Finally, the results of the two parts are combined, reducing the amount of computation and simultaneously improving the accuracy of the detection. The module structure is shown in Figure 11.

Figure 11. The architecture of SPPCSPC.

ELAN-F is similar to the ELAN structure in the Backbone but differs in that the number of outputs in the first branch is increased by summing each output section, allowing for
more efficient learning and convergence in a deeper network structure. The ELAN-F structure is shown in Figure 12.

![Figure 12. The architecture of ELAN-F.](image)

2.4.3. Head

In this part, YOLOv7 selects the ‘IDetect’ detection head with three target scales: large, medium, and small. The Head is used as the classifier and regressor of the network, and three enhanced effective feature layers are obtained through the above three parts. The information inside is used for feature-point judgment to determine whether there is a target to correspond to a priori box in the feature point. The use of the RepConv module allows the structure of the model to change during training and inference, introducing the idea of a re-parameterized convolution structure, as in Figure 13. RepConv is divided into two parts. The first uses three branches at training time; the top branch is a 3×3 convolution for feature extraction, the second branch is a 1×1 convolution for feature smoothing, adding a residual structure of Identity if the input and output are of equal size, and finally fusing and summing these three parts. At the time of inference, there is only one 3×3 convolution, which is re-parameterized from the training module.

![Figure 13. The architecture of RepConv.](image)

3. Underwater-YCC Algorithm

In this section, the Underwater-YCC target detection algorithm is introduced. The main structure diagram of this algorithm is shown in Figure 14.

![Figure 14. Main structure diagram of Underwater-YCC.](image)
3. Underwater-YCC Algorithm

In this section, the Underwater-YCC target detection algorithm is introduced. The main structure diagram of this algorithm is shown in Figure 14.

![Figure 14. The architecture of Underwater-YCC.](image)

3.1. YOLOv7 with CBAM

In the field of target detection, there is no single rule for where the best results can be achieved by adding attention mechanisms, and the results vary from location to location. For YOLOv7, three different fusion methods have been chosen for the three modules Backbone, Neck, and Head. The first is to add the attention mechanism to the Backbone section, which is part of the network where the features are extracted. The fusion of attention at this location can help the network to extract more effective information and locate fine-grained features more easily, thus improving the overall performance of the network. The second method is to add the attention mechanism to the Neck part of the network, where the features are integrated and extracted. When fusing information at different scales, adding the attention mechanism can help the network to fuse more valuable information into the features to refine the features. The last approach is to add the attention mechanism to the Head section, which is for feature classification as well as regression prediction, and to add the attention mechanism before the three different scales of features in and out, to perform attention reconstruction on the feature map and ultimately improve the network performance. The three attention mechanisms are added as shown in Figure 15.
The introduction of the transformer has given a huge boost to the field of computer vision, demonstrating powerful performance in areas such as image segmentation and target detection. More and more researchers are proposing the encoding of spatial features by convolution, and Conv2Former is one of the most efficient methods for encoding spatial features using convolution. The structure of Conv2Former [25] is shown in Figure 16, which is a transformer-style convolutional network with a pyramidal structure and a different number of convolutional blocks in each of the four stages. Each stage has a different feature map resolution, and a patch-embedding block is used in between two consecutive stages to reduce the resolution. The core of the method lies in the convolutional modulation operation, as shown in Figure 17, using only deep convolutional features as weights to modulate the representation, combined with Hadamard product to simplify the self-attentive mechanism and make more efficient use of large kernel convolution. Inspired by TPH-YOLOv5 [26], Conv2Former replaces the ELAN-F convolution block in the Neck of the original YOLOv7. Compared with the original structure, Conv2Former can better capture the global information and contextual semantic information of the network, and thus obtain rich features for fusion operation, which enables the network performance to be improved.

**Figure 15.** Left: Incorporate an attention mechanism in the Backbone. Middle: Incorporate an attention mechanism in the Neck. Right: Incorporate an attention mechanism in the Head.

**3.2. Neck Improvement Based on Conv2Former**

Figure 16. Overall architecture of Conv2Former.
3.3. Introduction of Wise-IoU Bounding Box Loss Function

In the field of target detection, the setting of the bounding box loss function directly affects the accuracy of the target detection result. The bounding box loss function is used to optimize the error between the position of the detected object and the real object so that the output prediction box is infinitely close to the real box. As the scenes and datasets faced in underwater practical work are of poor quality, we propose the use of Wise-IoU as the bounding box loss function, thus balancing the results of the model-trained images of varying quality to obtain a more accurate detection result. Wise-IoU [27] is a category weight introduced on top of the traditional IoU to minimize the difference between categories, thus reducing the impact on detection results.

The bounding box loss function is used to optimize the error between the position of the detected object and the real object so that the output prediction box is infinitely close to the real box. As the scenes and datasets faced in underwater practical work are of poor quality, we propose the use of Wise-IoU as the bounding box loss function, thus balancing the results of the model-trained images of varying quality to obtain a more accurate detection result. Wise-IoU [27] is a category weight introduced on top of the traditional IoU to minimize the difference between categories, thus reducing the impact on detection results. That is, a weight is assigned to each category and then the overlap between different categories is weighted using different weights in the calculation of IoU to obtain a more accurate evaluation result. Wise-IoUv1 with a two-level attention mechanism is first constructed based on the distance metric with the following equation:

$$L_{\text{Wise-IoUv1}} = R_{\text{Wise-IoU}} L_{\text{IoU}}$$

(5)

$$R_{\text{Wise-IoU}} = \exp \left( \frac{(x - x_{gt})^2 + (y - y_{gt})^2}{(W_g^2 + H_g^2)^r} \right)$$

(6)

An anchor box is represented by $$\bar{B} = [xywh]$$, where the value represents the center coordinates and size of the corresponding bounding box, and $$B_{gt} = [x_{gt} y_{gt} w_{gt} h_{gt}]$$ refers to the corresponding value of the target box. $$W_g$$ and $$H_g$$ are the minimum dimensions of the bounding box, $$R_{\text{Wise-IoU}}$$ can significantly amplify the IoU Loss of an ordinary quality anchor box, and $$L_{\text{IoU}}$$ can reduce $$R_{\text{Wise-IoU}}$$ of a high-quality anchor box. The method used in this paper applies Wise-IoU with $$\beta$$ on top of Wise-IoUv1. The outlier $$\beta$$ is used to describe the quality of anchor frames, with a smaller outlier representing a higher-quality anchor frame. A smaller gradient gain is assigned to anchor frames with a two-level allocation strategy and improve the positioning accuracy of the model.

$$L_{\text{Wise-IoU}} = r L_{\text{Wise-IoUv1}}, r = \frac{\beta}{\delta^2 - \delta}$$

(8)

The Wise-IoU used is defined as follows: $$\delta$$ makes $$r = 1$$ when $$\beta = \delta$$. The anchor box will have the highest gradient when the outlier is equal to a fixed value. According to Equation (7), the criteria for dividing the anchor box are dynamic, so Wise-IoU can use the best gradient gain allocation strategy and improve the positioning accuracy of the model.

![Figure 17. Left: Transformer; Right: Convolutional modulation.](image-url)
4. Experiments

4.1. Experimental Platform

The experimental environment of this paper is shown in Table 1.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>Intel(R) Core(TM) i9-10920X</td>
</tr>
<tr>
<td>GPU</td>
<td>NVIDIA GeForce RTX 3090</td>
</tr>
<tr>
<td>Operating system</td>
<td>Windows10</td>
</tr>
<tr>
<td>Frame</td>
<td>Pytorch1.7</td>
</tr>
<tr>
<td>CUDA</td>
<td>11.7</td>
</tr>
<tr>
<td>Batch Size</td>
<td>16</td>
</tr>
<tr>
<td>Epochs</td>
<td>300</td>
</tr>
<tr>
<td>Image Size</td>
<td>640 * 640</td>
</tr>
</tbody>
</table>

4.2. Evaluation Metrics

In this paper, the metric’s precision, recall, F1 score, and mAP are selected to evaluate the performance of the model. If the predicted value is the same as the true value, the predicted value is a positive sample, denoted TP. If the predicted value is a negative sample, it is denoted TN. If they are not the same, and the predicted value is a positive sample, it is denoted FP, and if the predicted value is a negative sample, it is denoted FN. The recall, precision and F1 score are calculated as follows:

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (9)
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (10)
\]

\[
F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (11)
\]

AP is the average of the precision values on the PR curve, obtained using different combinations of precision and recall points to calculate the area under the curve. mAP is the mean average precision; these metrics can be expressed as:

\[
AP = \int_0^1 P(R) dR \quad (12)
\]

\[
mAP = \frac{1}{\text{class\_num}} \int_0^1 P(R) dR \quad (13)
\]

4.3. Experimental Results and Analysis

The results in this section are obtained experimentally on the URPC2020 dataset. The mislabeled images in this dataset are re-labeled, the overly blurred images are filtered out, and the final experimental results are obtained on the optimized dataset.

4.3.1. Data Augmentation

Experiments were conducted using different data enhancement methods on the original structure of YOLOv7. From Table 2, the mAP of the model training results was only 64.59% when no data enhancement method was used, which increased by 4.91% and 17.38% after training with mixup and mosaic, respectively, and by 21.08% when the two enhancement methods were used together. The experimental results show that both data augmentation methods can help train the model well, and the use of both can greatly improve the detection accuracy of the model.
Table 2. Data Augmentation.

<table>
<thead>
<tr>
<th>Mixup</th>
<th>Mosaic</th>
<th>Precision</th>
<th>Recall</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>×</td>
<td>×</td>
<td>66.93%</td>
<td>60.63%</td>
<td>64.59%</td>
</tr>
<tr>
<td>✓</td>
<td>×</td>
<td>73.34%</td>
<td>64.08%</td>
<td>69.50%</td>
</tr>
<tr>
<td>×</td>
<td>✓</td>
<td>81.82%</td>
<td>75.09%</td>
<td>81.91%</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>84.21%</td>
<td>80.97%</td>
<td>85.67%</td>
</tr>
</tbody>
</table>

4.3.2. Fusion Attention Mechanism Comparison Test

The model and attention mechanism were optimally combined by adding the attention mechanism at different locations in YOLOv7, and CBAM was added to the Backbone, Neck, and Head parts of the network, respectively. Table 3 shows the experimental results. The addition of CBAM to the network improved the recognition accuracy of the network, with the best result being 86.68% at the Neck; both accuracy and recall were higher than the original model. The results show that CBAM does not work in all parts of the network. In the Head part, due to the deeper model, the underlying semantic information has been lost, and it is difficult to obtain results with fewer features for further attention weighting, so many metrics have decreased. The best embedding results are obtained in the Neck part, where the attentional weighting of the feature maps of different dimensions is more effective at obtaining fine-grained semantic information. This helps the network to grasp the detection target, and thus obtain the most significant effect.

Table 3. Fusion Attention Mechanism.

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOLOv7</td>
<td>84.21%</td>
<td>80.97%</td>
<td>85.67%</td>
</tr>
<tr>
<td>v7_Backbone</td>
<td>84.00%</td>
<td>81.57%</td>
<td>86.11%</td>
</tr>
<tr>
<td>v7_Neck</td>
<td>84.90%</td>
<td>80.67%</td>
<td>86.68%</td>
</tr>
<tr>
<td>v7_Head</td>
<td>83.15%</td>
<td>81.05%</td>
<td>85.61%</td>
</tr>
</tbody>
</table>

4.3.3. Ablation Experiments

In order to verify the effectiveness of each improved method for underwater target detection, the effect of different modules on detection results is analyzed by ablation experiments. Among them, YOLOv7_A adds CBAM to the Neck, YOLOv7_B uses Conv2Former to improve the Neck, YOLOv7_C uses Wise-IoU, YOLOv7_D uses both CBAM and Wise-IoU, and YOLOv7_E uses both Conv2Former and Wise-IoU. Underwater-YCC is the underwater target detection method proposed in this paper.

From Table 4, we can see that the experimental results obtained for each of the modular methods used are improved compared to the original YOLOv7, indicating that all reinforcement methods used in this paper are effective and can all be used to improve underwater detection activities. (1) Analyzing the results of the three single methods in experiments (a–c) shows that the addition of each optimization method is improved compared to YOLOv7, where the addition of Conv2Former has improved the mAP of the network by 0.85%. This means that the Conv2Former module can capture the global information of the network well and retain the semantic information. The introduction of CBAM gives the network the ability to acquire more valuable features for fusion. The 0.88% improvement using Wise-IoU means that using this method allows the network to focus more on effective features and have better weight selection for images of different quality. (2) The results of experiments (d,e) show that combining Wise-IoU with CBAM and Conv2Former, respectively, improves 1.17% and 1.26%, compared to YOLOv7, indicating that this bounding box loss function is effective after adding the optimization method. (3) After summarizing the above optimization methods, this paper proposes an optimization algorithm for Underwater-YCC, which adds CBAM while using Conv2Former for Neck feature fusion, and lastly uses Wise-IoU for bounding box loss regression. This model improved the mAP
by 1.49% compared to the original YOLOv7. The results show that the Underwater-YCC method can perform high-quality detection in complex underwater environments.

Table 4. Ablation Experiments.

<table>
<thead>
<tr>
<th>Model</th>
<th>CBAM</th>
<th>Conv2Former</th>
<th>Wise-IoU</th>
<th>Precision</th>
<th>Recall</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOLOv7</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>84.21%</td>
<td>80.97%</td>
<td>85.67%</td>
</tr>
<tr>
<td>(a) YOLOv7_A</td>
<td>√</td>
<td>×</td>
<td>×</td>
<td>84.90%</td>
<td>80.67%</td>
<td>86.68%</td>
</tr>
<tr>
<td>(b) YOLOv7_B</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>83.97%</td>
<td>81.84%</td>
<td>86.52%</td>
</tr>
<tr>
<td>(c) YOLOv7_C</td>
<td>×</td>
<td>×</td>
<td>√</td>
<td>82.53%</td>
<td>82.01%</td>
<td>86.55%</td>
</tr>
<tr>
<td>(d) YOLOv7_D</td>
<td>√</td>
<td>×</td>
<td>√</td>
<td>85.24%</td>
<td>79.84%</td>
<td>86.84%</td>
</tr>
<tr>
<td>(e) YOLOv7_E</td>
<td>×</td>
<td>√</td>
<td>√</td>
<td>84.26%</td>
<td>81.06%</td>
<td>86.93%</td>
</tr>
<tr>
<td>Underwater-YCC</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>84.64%</td>
<td>81.38%</td>
<td>87.16%</td>
</tr>
</tbody>
</table>

Figure 18 depicts the test results of Underwater-YCC compared with YOLOv7. Among them, Figure 18a is the detection result of YOLOv7 and Figure 18b is the detection result of Underwater-YCC. From the figures, we can get that our proposed model can detect more targets compared with the original model and has better results for the detection of complex underwater environments.

4.3.4. Target Detection Network Comparison Experiment Results

Table 5 compares the results of Underwater-YCC with classical target detection algorithms, such as Faster-RCNN [28], YOLOv3, YOLOv5s, YOLOv6 [29], and YOLOv7-Tiny. It can be seen from the results that although the detection time increases slightly due to the
complex structure of the model, Underwater-YCC has higher detection accuracy and is more adaptable to the complex underwater environment.

Table 5. Compare with classical target detection algorithms.

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>mAP</th>
<th>F1 Score</th>
<th>FPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster-RCNN</td>
<td>38.3%</td>
<td>55.26%</td>
<td>62.1%</td>
<td>45.24</td>
<td>16</td>
</tr>
<tr>
<td>YOLOv3</td>
<td>76.6%</td>
<td>64.3%</td>
<td>80.9%</td>
<td>69.91</td>
<td></td>
</tr>
<tr>
<td>YOLOv5s</td>
<td>83.34%</td>
<td>78.96%</td>
<td>83.88%</td>
<td>81.09</td>
<td>46.51</td>
</tr>
<tr>
<td>YOLOv6</td>
<td>82.1%</td>
<td>63.6%</td>
<td>82.09%</td>
<td>71.67</td>
<td></td>
</tr>
<tr>
<td>YOLOv7-Tiny</td>
<td>81.44%</td>
<td>79.14%</td>
<td>84.21%</td>
<td>80.27</td>
<td>32.89</td>
</tr>
<tr>
<td>YOLOv7</td>
<td>84.21%</td>
<td>80.97%</td>
<td>85.67%</td>
<td>82.55</td>
<td>26.42</td>
</tr>
<tr>
<td>Underwater-YCC</td>
<td>84.64%</td>
<td>81.38%</td>
<td>87.16%</td>
<td>82.97</td>
<td>21.17</td>
</tr>
</tbody>
</table>

5. Conclusions

In this study, we addressed the challenges of false and missed detection caused by blurred underwater images and the small size of underwater creatures. To tackle these issues, we proposed an underwater target detection algorithm called Underwater-YCC based on YOLOv7. We tested our algorithm on the URPC2020 dataset, which includes underwater images of echinus, holothurian, scallop, and starfish categories.

Our proposed algorithm leverages various techniques to improve detection accuracy. Firstly, we reorganized and labeled the dataset to better suit our needs. Secondly, we embedded the attention mechanism in the Neck part of YOLOv7 to improve the detection ability of the model. Thirdly, we used Conv2Former to enable the network to obtain features that are more valuable and fuse them efficiently. Lastly, we used Wise-IoU for bounding box regression calculation to effectively avoid the drawbacks caused by the large sample gap.

Experimental results demonstrate that the Underwater-YCC algorithm can achieve improved detection accuracy under the same dataset. Our approach also exhibits robustness in the case of blurring and color bias. However, there is still ample room for improving the whole network structure, and the real-time and lightweight aspects of the underwater target detection technology need to be studied further. The proposed algorithm is promising and may serve as a starting point for future research in the field of underwater target detection.

Author Contributions: Conceptualization, X.C. and M.Y.; Formal analysis, Q.Y., H.Y. and H.W.; Funding acquisition, X.C. and H.W.; Investigation, Q.Y.; Methodology, X.C. and M.Y.; Resources, Q.Y. and H.Y.; Software, M.Y.; Validation, M.Y.; Writing—original draft, X.C. and M.Y.; Writing—review & editing, X.C., M.Y. and H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key project of National Natural Science Foundation of China, grant number 62031021, and Natural Science Foundation of Shaanxi Province, China, grant number 20JK0532.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

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

References


17. 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]


20. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. J. Big Data 2019, 6, 60. [CrossRef]


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