Underwater Small Target Detection Based on YOLOX Combined with MobileViT and Double Coordinate Attention

Yan Sun 1, Wenxi Zheng 1,*, Xue Du 2 and Zheping Yan 2

1 College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China
2 College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China; duxue@hrbeu.edu.cn (X.D.); yanzheping@hrbeu.edu.cn (Z.Y.)
* Correspondence: zhengwenxi@hrbeu.edu.cn

Abstract: The underwater imaging environment is complex, and the application of conventional target detection algorithms to the underwater environment has yet to provide satisfactory results. Therefore, underwater optical image target detection remains one of the most challenging tasks involved with neighborhood-based techniques in the field of computer vision. Small underwater targets, dispersion, and sources of distortion (such as sediment and particles) often render neighborhood-based techniques insufficient, as existing target detection algorithms primarily focus on improving detection accuracy and enhancing algorithm complexity and computing power. However, excessive extraction of deep-level features leads to the loss of small targets and decrease in detection accuracy. Moreover, most underwater optical image target detection is performed by underwater unmanned platforms, which have a high demand of algorithm lightweight requirements due to the limited computing power of the underwater unmanned platform with the mobile vision processing platform. In order to meet the lightweight requirements of the underwater unmanned platform without affecting the detection accuracy of the target, we propose an underwater target detection model based on mobile vision transformer (MobileViT) and YOLOX, and we design a new coordinate attention (CA) mechanism named a double CA (DCA) mechanism. This model utilizes MobileViT as the algorithm backbone network, improving the global feature extraction ability of the algorithm and reducing the amount of algorithm parameters. The double CA (DCA) mechanism can improve the extraction of shallow features as well as the detection accuracy, even for difficult targets, using a minimum of parameters. Research validated in the Underwater Robot Professional Contest 2020 (URPC2020) dataset revealed that this method has an average accuracy rate of 72.00%. In addition, YOLOX’s ability to compress the model parameters by 49.6% efficiently achieves a balance between underwater optical image detection accuracy and parameter quantity. Compared with the existing algorithm, the proposed algorithm can carry on the underwater unmanned platform better.

Keywords: YOLOX; underwater target detection; MobileViT; coordinate attention

1. Introduction

The rapid advancement of science and technology has led to a shift in human behavior and spurred an interest in the exploitation of aquatic resources at the expense of land-based ones [1]. Since 70% of the planet is covered by water, there are plenty of underwater resources to drive future advances in science and technology, and, as the forerunner of subsurface resource utilization, underwater exploration has had ample time to make significant advances. Several researchers have devoted themselves to the study of underwater optical image detection, a cornerstone of modern underwater detection that has been used successfully in many areas of ocean exploration. However, underwater detection is limited by several issues that do not impede land detection. The complexity of the underwater imaging environment causes problems, such as location dispersion, color deviation, and
blur, to name just a few. Furthermore, determining the target volume of an underwater object presents additional challenges. These factors make conventional target detection techniques incapable of meeting the demands of the modern industry. In addition to these issues, the limited use of underwater detection and underwater unmanned platforms transport has resulted in storage limitations for underwater equipment and vehicles, and the typically large models that land algorithms are based on are difficult to transport in underwater environments. Therefore, the development of an underwater optical image detection method with low parameters and high accuracy that is suited to the needs of underwater unmanned platforms is essential.

Existing target detection algorithms are mainly divided into two types: two-stage [2–4] and one-stage [5–8]. The former has stronger detection accuracy but a complex structure, while the latter has lower accuracy but a lightweight structure. In order to ensure the carrying on the underwater unmanned platform, we chose YOLOX [9] in the first-order algorithm. On the basis of inheriting CSPDarknet53 [10] and the feature pyramid network (FPN) [11] of YOLO series algorithms, YOLOX applies frameless detection in the YOLO algorithm for the first time, which reduces the computational complexity of YOLOX. However, due to the complex feature extraction of CSPDarknet, it is not friendly to the detection of underwater targets. In addition, the underwater unmanned platform has higher requirements for the storage of the algorithm, but the existing lightweight algorithm has insufficient feature extraction ability. Therefore, we chose the MobileViT [12] lightweight model as the backbone network of the algorithm. In addition, in order to extract the shallow information of the target better, we proposed a new attention mechanism DCA based on CA [13] attention and applied it to YOLOX, so that the algorithm can obtain higher accuracy. Experiments show that the accuracy of URPC2020 data is up to 72.00% and the number of parameters is reduced by 49.6%.

With this background in mind, we present the main contributions of this paper:

1. Within the mainstream of target detection methods, we chose YOLOX algorithms as the basic structure. By using MobileViT as the backbone network in YOLOX, we further improved the global feature extraction ability of the algorithm while reducing the number of parameters.

2. To address problems posed by underwater targets characterized by small volumes, scattered distributions, and blurred imaging, we designed a DCA mechanism based on prior CA mechanisms. By improving the shallow feature extraction ability of the algorithm model, we enhanced its ability to extract the data of difficult targets.

3. The results of our evaluation of the URPC2020 dataset show that our network model has better accuracy compared with the baseline method while reducing the number of parameters. Therefore, our method is not only feasible but also superior to the original baseline method.

2. Related Work

2.1. Object Detection

Researchers in the field of underwater target detection have applied convolutional neural networks (CNNs) extensively [14], and the existing target detection algorithms based on CNNs are mainly divided into two types: one-stage algorithms and two-stage algorithms. Two-stage algorithms, on the other hand, extract a series of candidate regions and classify them for target detection. Two-stage algorithms include the R-CNN [2], Fast R-CNN [3], and FasterR-CNN [4], among others. Two-stage algorithms have high accuracy, but their detection efficiency is lower than that of one-stage algorithms. One-stage algorithms use a first-order network to complete classification and location tasks, greatly improving detection efficiency and achieving a good balance between accuracy and algorithm volume. Some first-stage algorithms include the Single-Shot MultiBox Detector (SSD) [15] and the YOLO (You Only Look Once) [5–8] series. Although the YOLO series algorithm achieves a good balance in accuracy and complexity, the YOLO series algorithm has a poor detection effect on small targets and low recall rate, which affects the application
of the YOLO algorithm underwater. Therefore, researchers have carried out a lot of studies based on YOLO series algorithms.

Chen et al. [16] proposed an underwater target recognition network based on improved YOLOv4. Lei et al. [17] applied YOLOv5 in underwater target detection. These tasks integrate the YOLO series into underwater target detection, Chen et al. [18]. Proposed an underwater target detection lightweight algorithm based on multi-scale feature fusion but do not take into account the hardware limitations of underwater detection.

Underwater targets are frequently dispersed, resulting in the loss of small targets, and the deep extraction of features often produces images that are blurred [19]. Therefore, there is a need to improve algorithms’ abilities to extract shallow information, the low-level features and patterns typically captured by the initial layers of a neural network. Our proposed model extracts shallow information by utilizing an attention mechanism during the shallow information extraction stage.

In order to solve the problems surrounding feature extraction, we propose the introduction of a transformer [20]. Although transformers are still in their early stages, transformer-based models have already achieved excellent results in the field of natural language processing. In 2020, researchers applied the transformer model to computer vision [21] domain tasks and achieved promising results. Algorithms built on transformer-based models are better able to focus on feature extraction and, therefore, have stronger global feature extraction capabilities. Lei et al. [17] proposed YOLOv5 combined with a Swin Transformer. Chen et al. [22] proposed a lightweight underwater target detection algorithm based on dynamic sampling transformer and knowledge-distillation optimization, but they ignored the cost of the transformer.

Researchers are also starting to incorporate lightweight transformers into the algorithms they use for underwater detection. In 2021, MobileVIT combined a transformer with a lightweight algorithm for the first time, which can greatly reduce the amount of calculation, while guaranteeing the feature extraction capability, and can meet the needs of underwater unmanned platforms. Thus, we chose MobileVIT [12] to combined with YOLOX [9].

2.2. Lightweight Network and Attention

Considering the storage limitations of underwater unmanned platforms, underwater target detection algorithms normally face additional challenges due to large amount of algorithm parameters, which limits their speed and efficiency. To address this issue, lightweight algorithm research for target detection has attracted the attention of researchers in recent years and they have sought to develop algorithms optimized for resource-constrained devices. MobileNet (v1/v2) [23,24], SqueezeNet [25], ShuffleNet [26], GhostNet [27], and other lightweight CNN architectures have all been developed to provide deep learning algorithms that can extract features more efficiently by improving their convolution methods. Yeh et al. [28] proposed a lightweight deep neural network for joint learning of underwater object detection and color conversion. Wang et al. [29] applied a novel attention-based lightweight network for multiscale object detection in underwater images. The downside is that the use of lightweight methods inevitably leads to reductions in detection accuracy.

In underwater target detection, attention mechanisms are frequently used in feature extraction, and, in mobile networks, attention mechanisms have proven their usefulness in computer vision through their ability to achieve efficient feature extraction at a relatively low cost. The squeeze-and-excitation network (SENet) [30], convolutional block attention module (CBAM) [31], and CA mechanism [13] are among the definitive attention mechanisms noted for their high efficiency. SENet compresses and maps 2D features to prioritize informative channels; CBAM further improves on spatial information coding by applying a large-size kernel to the feature map and using convolutional layers. Zhang et al. [32] proposed underwater object detection based on YOLOv4 and multi-scale attentional feature fusion. Li et al. [33] applied YOLOv4 combined channel attention to detect underwater
biochemistry. In order to enhance the feature extraction capability of the algorithm, we designed a new attention mechanism named DCA. DCA can integrate feature information extracted by CNN and the transformer to further enhance the shallow information extraction of the algorithm.

The rest of this paper is organized as follows: in Section 3, we introduce the YOLOX algorithm, MobileViT algorithm, and CA attention mechanism based on which the algorithm in this paper is based; in Section 4, we introduce the new DCA attention mechanism proposed in this paper and the improved algorithm combining the MobileViT and YOLOX algorithm; in Section 5, we verify the effectiveness of the algorithm through experiments; and in the fifth chapter, we draw a conclusion. The flow chart of this paper is shown in Figure 1.

**Figure 1.** The flow chart of algorithm structure.

### 3. Basic Structure

#### 3.1. YOLOX Structure

YOLOX improves on prior YOLO object detection algorithms, which are a series of single-shot detectors developed by Megvii Technology in 2021. A key innovation of YOLOX is the compartmentalization of tasks performed by separate components, such as the image pre-processor, the backbone network, the neck network, and the predictive head network. YOLOX uses mosaic data enhancement during the image pre-processing stage and selects four images from the dataset for stitching and testing, which can enrich image backgrounds. In the backbone network, YOLOX uses CSPDarknet CNNs for feature extraction, consistent with previous generations of YOLO algorithms. FPN and PANNet are used in the neck network to combine features from different layers, allowing shallow information to guide the deep information, thereby retaining the position information of the input image. A major innovation of YOLOX is its decoupled head architecture; the head network makes predictions about objects in the input image—such as target category, background, and coordinate information—before relegating the classification and regression tasks into separate modules. This improves the expression ability and accuracy of the algorithm and accelerates convergence. In addition, YOLOX optimizes the number of parameters and further improves the accuracy of the algorithm using anchor-free [34] detection and SimOTA [35] tag allocation. Despite its excellent detection accuracy, YOLOX operates at a low cost; therefore, we chose YOLOX as our baseline method. In order to extract special features of underwater targets, we optimized the original YOLOX by further improving the model’s detection accuracy and reducing the number of parameters. The YOLOX structure is shown in Figure 2.
3.2. MobileViT Structure

MobileViT is a computer vision model that combines mobile-friendly CNNs with vision transformers. CNNs focus on the extraction of local information, ignoring correlations within this information, and their use of excessive convolution leads to the loss of key information from the target. Compared with CNNs, transformer methods perform better in global feature extraction, and transformers are also better able to identify correlations between adjacent positions, improving how the shallow information of an image is saved. However, because the transformer is a heavyweight model lacking the inductive bias that would allow it to migrate directly to target detection, it often results in algorithms that operate poorly. MobileViT, on the other hand, combines CNN and transformer layers, resulting in a model featuring the efficient and lightweight aspects of CNNs as well as the strong overall vision capacity of the transformer method. The two core components of MobileViT are MobileViTBlock and MobileNetV2Block. MobileNetV2Block is the inverted residual block component in MobileNetV2. The Mobilenet v2 block is shown in Figure 3.

This component enables an algorithm to retain its CNN while significantly reducing the amount of calculation and number of parameters required, while additionally avoiding the need for the extensive transformer operations typically performed by MobileViT. MobileNetBlock, a building block in MobileNet, first adjusts the number of channels through \(3 \times 3\) convolution and \(1 \times 1\) convolution, and then extracts global features through the unfold, transformer, and fold technique. After the channel is adjusted via \(1 \times 1\) convolution, it is concatenated with the original special diagnosis map by shortcut, and then feature fusion is performed by \(3 \times 3\) convolution. The MobileViT block is shown in Figure 4.
Figure 4. The network structure of MobileVIT block.

3.3. Coordinate Attention

Attention mechanisms have been proven to significantly improve the performance of neural networks. However, due to the large amount of computation required by self-attention modules (also known as self-attention mechanisms), they can only be used in large models and are not suitable for mobile networks. Most of the attention mechanisms applicable to mobile networks are based on global pooling, which reduces the spatial dimensions of a feature map to a single value per channel—ignoring the location information of the features. This lack of position information affects the structure of the object captured by target detection. In addition, CNNs can only capture local relationships; our proposed coordination attention mechanism solves this problem by decomposing channel attention into two one-dimensional feature coding processes, which are characterized in two directions. This allows for the capture of remote dependencies in one direction while retaining accurate location information in the other direction. The two directional feature maps are then combined with the input feature maps to enhance the direction and position information of the target. This method allows an algorithm to focus on the global area of a network at a low computational cost, while the use of two parallel one-dimensional feature decoding can reduce the feature information loss caused by global pooling.

We use spatial extents of pooling kernels \((H, 1)\) or \((1, W)\) to encode each channel along the horizontal coordinate and the vertical coordinate separately. Thus, the output of the \(c\)-th channel at height \(h\) can be formulated as:

\[
Z_c^w (w) = \frac{1}{W} \sum_{0 \leq i < W} x_c (h, i)
\]  

Similarly, the output of the \(c\)-th channel at width \(w\) can be formulated as:

\[
Z_c^w (w) = \frac{1}{H} \sum_{0 \leq j < H} x_c (j, w)
\]  

Specifically, Equations (1) and (2) produce the aggregated feature maps, and then we start to concatenate and then send them to a shared \(1 \times 1\) convolutional transformation function \(F_1\), which can be formulated as:

\[
f = \delta \left( F_1 \left( \left[ z^h, z^w \right] \right) \right)
\]  

where \(\left[ \cdots \right]\) denotes the splicing operations along the spatial dimension, \(\delta\) denotes the non-linear activation function, and \(f \in \mathbb{R}^{C/r \times (H+r \times W)}\) is the intermediate feature map encoding spatial information in both horizontal and vertical directions. In addition, \(r\) is the reduction ratio. We then split \(f\) along the spatial dimension into two separate tensors, \(f^h \in \mathbb{R}^{C/r \times W}\) and \(f^w \in \mathbb{R}^{C/r \times H}\). \(f_h\) and \(f_w\) are \(1 \times 1\) convolutional transformations which...
are utilized to separately transform $f^h$ and $f^w$ to tensors with the same channel number as the input, which can be formulated as:

$$g^h = \sigma(F_h(f^h))$$  \hspace{1cm} (4)  

$$g^w = \sigma(F_w(f^w))$$  \hspace{1cm} (5)

where $\sigma$ denote the sigmoid function. To satisfy the demand of the lightweight model, we frequently reduce the channel number $f$ with an appropriate reduction ratio $r$. The outputs $g^h$ and $g^w$ are expanded and used as attention weights separately. Finally, Figure 5 shows the structure of the CA mechanism, and the output of the coordinate attention block can be formulated as:

$$y_c(i, j) = x_c(i, j) \times g^h_c(i) \times g^w_c(j)$$  \hspace{1cm} (6)

Figure 5. The network structure of coordinate attention.

4. Proposed Structure

4.1. Double Coordinate Attention

Underwater small target detection methods require improvements in shallow information feature extraction. To this end, we found that shallow feature extraction can be further strengthened by utilizing a CA mechanism and fusing the feature information in the backbone network before and after the block. Additionally, we aggregated the features in different directions, so that the algorithm can focus more attention on the location of shallow information. After obtaining the weight in each direction, it is combined with the input from both sides, allowing the shallow information of the output feature map to be better expressed. With this improved backbone network structure, we confirmed
that MobileViT block does not change the structure of the input features when features are fused before or after this group of layers. Thus, we designed proposed coordinate attention named double coordinate attention (DCA). We tested this in consideration of how global features and local features are emphasized during feature extraction, and the results suggest distinct advantages for detecting the location of small targets.

We encoded the two input feature graphs $Z_n$ and $Z_v$ along two directions, respectively, and carried out feature convolution feature fusion on the feature codes in the same direction. Therefore, the horizontal output after convolution fusion $Z^h$ can be formulated as:

$$Z^h = Z^h_n(h) \otimes Z^h_v(h)$$  \hspace{1cm} (7)$$

Similarly, the vertical output $Z^w$ can be formulated as:

$$Z^w = Z^w_n(w) \otimes Z^w_v(w)$$  \hspace{1cm} (8)$$

where $\otimes$ represents the convolution operation. $n$ and $v$ represent the feature graphs of different inputs, respectively. Similar to the CA mechanism, we convolved the two input feature graphs and multiplied them with the attention weights of the two directions $g^h$ and $g^w$ to obtain the output $Y$, which can be formulated as:

$$y_c(i, j) = (x_n(i, j) \otimes x_v(i, j)) \times g^h_c(i) \times g^w_c(j)$$  \hspace{1cm} (9)$$

By combining the two input weights and inputting the special detection map with attention into FPN, the multi-feature extraction can be further guided. The DCA attention mechanism extracted in this paper can not only retain the acquisition of global features in the transformer, but also take into account the capture of local features in CNN and enhance the feature extraction of shallow information. This can better satisfy the needs of underwater target detection. Figure 6 shows the structure of the DCA mechanism.

Figure 6. The network structure of double coordinate attention.
4.2. Improved Network Structure

Because YOLOX has advantages in image enhancement, target classification, and label classification in the YOLO series of algorithms, we optimized the backbone network and feature extraction. YOLOX follows the backbone network of the previous generation of YOLO algorithm, CSPDarknet53, which has excellent and different performance in feature extraction. However, because the extraction of shallow information for underwater small targets is very important, the multi-layer convolution of CSPDarknet will cause serious loss of shallow information, make the detection of small targets more difficult, and the multi-layer convolution layer will make the algorithm cumbersome. Therefore, we chose MobileVIT as the backbone network of the algorithm for small target feature extraction. MobileVIT combines the lightweight algorithm MobileNet and vision transformer. It not only retains the lightweight and efficient features of CNN, but also reflects the feature that the transformer focuses on global features. For the support of shallow information to small targets, we added the attention mechanism DCA when outputting the first two features, which can greatly enhance the extraction of shallow information in the algorithm and enhance the detection accuracy of underwater small targets. Figure 7 shows the structure of the proposed YOLOX.

![Figure 7. The network structure of improved YOLOX.](image)

5. Experiment and Analysis

5.1. Datasets

This experiment uses the URPC2020 dataset for validation experiments, which includes five categories: starfish, scallops, waterweeds, echinus, and holothurian. It also included 5543 training images, of which aquatic plants were officially recognized as a negligible target, and it only contains 82 targets. However, in order to verify the algorithm’s ability to detect small underwater targets, we still included aquatic plants in the detection category, and this paper ultimately uses 5543 images and five categories and, according to the ratio of 9:1, it is randomly divided into training set and validation set.

5.2. Experimental Environment

The experimental environment is shown in Table 1. The hardware used for this experiment was a Ryzen 7 3700x; NVIDIA RTX 3090 graphics card; Ubuntu 18.04
operating system; CUDA version 11.6; PyTorch version 1.11.0; and Python environment 3.9.12.

Table 1. The environment for the experiments.

<table>
<thead>
<tr>
<th>Environment</th>
<th>Versions or Model Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>Ryzen 7 3700x</td>
</tr>
<tr>
<td>GPU</td>
<td>NVIDIA RTX 3090</td>
</tr>
<tr>
<td>OS</td>
<td>Ubuntu 16.04</td>
</tr>
<tr>
<td>CUDA</td>
<td>11.6</td>
</tr>
<tr>
<td>Python</td>
<td>V1.11.0</td>
</tr>
<tr>
<td>Pytorch</td>
<td>V3.9.12</td>
</tr>
</tbody>
</table>

5.3. Parameter Settings

We validated the improved YOLOX algorithm proposed in this paper on the URPC2020 dataset. The input resolution was uniformly resized to 608 × 608. For the task, the original YOLOX-S model was adopted as the basis for experimental comparison. For fair comparison, we trained all models across 300 epochs by using SGD with weight decay of 0.0005. The initial learning rate was set to 0.001 and the batch size was set to 16. All other parameters were kept the same as YOLOX-S.

5.4. Results and Analysis

We used the URPC dataset to train and test the model. mAP was the standard to measure the accuracy of the model in target detection. Since the model size was also one of the evaluation criteria in this paper, we also took the parameter size, Flops, as one of the measurement criteria. In order to verify the effectiveness of the experiment, this paper uses the ablation experiment, in which Model 1 uses the main framework of the YOLOX algorithm, and replaces the backbone network of YOLOX with MobileViT. Among the models, Model 1 adopts the main framework of the YOLOX algorithm, while replacing the backbone network of YOLOX with MobileViT, and replacing the convolution in FPN with depthwise separable convolution. The rest of the network structure is consistent with YOLOX. Model 2 is based on Model 1 and adds CA attention mechanism to the two layers of shallow information output, with the rest remaining unchanged. Model 3 is based on Model 2 and replaces the shallowest layer’s CA attention mechanism with the DCA attention mechanism, with the rest remaining unchanged. Model 4 is based on Model 2 and replaces both instances of the CA attention mechanism with the DCA attention mechanism, with the rest remaining unchanged.

The results in Table 2 show that using MobileViT as the backbone network in this paper demonstrates the advantage of MobileViT in reducing the number of parameters and improving the accuracy of target detection. Furthermore, the added attention mechanism proves that the coordinate attention mechanism can less sacrifice the parameter and Flops, and the DCA attention mechanism proposed in this paper can also better control the number of parameters. Model 4, proposed in this paper, can meet the carrying demand of underwater unmanned platforms.

Table 2. Parameter of ablation experiment.

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>Parameter(M)</th>
<th>Flops</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Dw</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YOLOX</td>
<td>√</td>
<td>8.94</td>
<td>26.64</td>
</tr>
<tr>
<td>Model 1</td>
<td>√</td>
<td>4.37</td>
<td>24.88</td>
</tr>
<tr>
<td>Model 2</td>
<td>√</td>
<td>4.39</td>
<td>24.92</td>
</tr>
<tr>
<td>Model 3</td>
<td>√</td>
<td>4.42</td>
<td>25.18</td>
</tr>
<tr>
<td>Model 4</td>
<td>√</td>
<td>4.51</td>
<td>25.35</td>
</tr>
</tbody>
</table>
The results in Table 3 show that Model 1 proposed in this paper demonstrates the advantages of the transformer algorithm in feature extraction, and Model 4, compared with Model 2 and Model 3, proves that the proposed DCA attention mechanism can perform better in shallow information extraction. Furthermore, the water seeds category listed in this paper has a significant increase in mAP compared with the baseline algorithm, proving that the DCA attention mechanism has a better prospect for detecting difficult and small targets.

### Table 3. Results of ablation experiment on the URPC2020 dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>mAP</th>
<th>Holothurian</th>
<th>Echinus</th>
<th>Starfish</th>
<th>Scallop</th>
<th>Waterweeds</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOLOX</td>
<td>66.92</td>
<td>67.00</td>
<td>87.09</td>
<td>79.49</td>
<td>83.05</td>
<td>17.96</td>
</tr>
<tr>
<td>Model 1</td>
<td>68.69</td>
<td>71.64</td>
<td>87.13</td>
<td>80.59</td>
<td>83.23</td>
<td>20.87</td>
</tr>
<tr>
<td>Model 2</td>
<td>71.01</td>
<td>73.24</td>
<td>86.93</td>
<td>80.42</td>
<td>82.81</td>
<td>31.66</td>
</tr>
<tr>
<td>Model 3</td>
<td>70.75</td>
<td>75.23</td>
<td>87.19</td>
<td>80.29</td>
<td>82.84</td>
<td>28.22</td>
</tr>
<tr>
<td>Model 4</td>
<td>72.00</td>
<td>73.42</td>
<td>87.37</td>
<td>79.40</td>
<td>82.97</td>
<td>36.87</td>
</tr>
</tbody>
</table>

The results in Table 4 show that compared with existing algorithms, the proposed algorithm achieves the best balance in precision and parameter quantity, and proves that the proposed algorithm can be carried on the underwater unmanned platform. Figure 8 shows the results of the proposed method for the detection.

### Table 4. Result of different algorithms on the URPC2020 dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>mAP</th>
<th>Holothurian</th>
<th>Echinus</th>
<th>Starfish</th>
<th>Scallop</th>
<th>Parameter (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOLOv4</td>
<td>81.01</td>
<td>71.21</td>
<td>89.94</td>
<td>85.58</td>
<td>77.30</td>
<td>64.04</td>
</tr>
<tr>
<td>T-YOLOv4</td>
<td>68.69</td>
<td>54.09</td>
<td>80.43</td>
<td>77.94</td>
<td>58.87</td>
<td>5.96</td>
</tr>
<tr>
<td>YOLOX</td>
<td>79.16</td>
<td>67.00</td>
<td>87.09</td>
<td>79.40</td>
<td>83.05</td>
<td>8.94</td>
</tr>
<tr>
<td>Model 4</td>
<td>80.79</td>
<td>73.42</td>
<td>87.37</td>
<td>79.40</td>
<td>82.97</td>
<td>4.51</td>
</tr>
</tbody>
</table>

Note: Data YOLOv4 and T-YOLOv4 are quoted from the literature [32].

![Figure 8](image_url)  
**Figure 8.** The detection results of our method in URPC2020.
6. Discussion
6.1. Underwater Target Detection Combined with Transformer

CNN-based object detection algorithms have always been a research hotspot and have been considered the foundation of object detection algorithms for many years. However, in underwater environments, the lack of global features and target loss caused by multiple convolutions limit the development of object detection algorithms. Previously, transformers have been widely used in the field of natural language processing. It was not until 2021 that transformers entered the field of image processing and achieved good results. Therefore, this paper focuses on combining MobileViT with YOLOX for object detection. Experiments show that this method can improve the accuracy of object detection while reducing the number of parameters. Thus, the transformer’s ability to extract features from targets is more suitable for underwater imaging environments, reducing target loss and demonstrating better accuracy for detecting difficult targets. However, since transformers need to consider global features, the detection speed will face significant challenges. Therefore, the optimization of the network structure should also consider improving the detection speed. In addition, compared with the existing YOLO algorithm, CNN-based deep feature information extraction still has certain advantages in high-precision categories. Therefore, further experiments are needed to combine CNN with the transformer.

6.2. Challenges of Underwater Small Target Detection

Existing detection algorithms enhance object detection accuracy through data augmentation, multi-feature fusion, and attention mechanisms. Among them, the impact of feature maps at different scales in multi-feature fusion on small object detection varies. In the experiments, we added attention mechanisms at different scales to enhance feature extraction, but the experimental results show that the large scale has a much greater impact on shallow information extraction than the small scale, and adding attention at the small scale may even reduce algorithm accuracy. Therefore, for underwater small targets, there should be more improvements in feature fusion. In addition, attention mechanisms mostly act as removable components added to existing algorithms, and experimental results show that attention mechanisms perform well in underwater small object detection. However, attention mechanisms rely too heavily on the structure of existing algorithms, and the improvement of experimental results is uncertain. The DCA attention mechanism proposed in this paper is not universal in improving the accuracy of target detection, and some categories are difficult to be improved. Therefore, there is still considerable room for the development of attention mechanisms in the field of underwater object detection.

6.3. Future Research Focus on Underwater Small Target Detection

From the experimental results, we analyzed that the reason why sea urchins can obtain the best performance is that the color of the sea urchin is monotonous and has obvious color difference with the color of underwater imaging, and the underwater dataset is large. Compared with other targets with high accuracy, the sea cucumber has a similar color and smaller size to the underwater background. Moreover, compared with the YOLOX algorithm, the detection rate of sea-participating seagrass has been significantly improved after combining with MobileViT feature extraction. Therefore, we believe that shallow feature extraction is positive for the detection of small underwater targets. Therefore, in the detection of small underwater targets, it is necessary to strengthen the shallow features. In addition, since the detection category with color difference has higher accuracy, it may bring unexpected results to be added into the target detection algorithm, such as reducing underwater color bias and increasing contrast.

7. Conclusions

This paper proposes an underwater object detection model that provides a good balance between accuracy and memory. In our work process, in order to establish an underwater object detection algorithm that can be mounted on underwater unmanned
platforms, we chose YOLOX as the base algorithm and used MobileViT as the backbone network to replace YOLOX’s CSPDarknet53 for extracting global features from images. At the same time, depthwise separable convolution is used in the neck for reconstruction to control the number of parameters in the algorithm. We also improved the DCA attention mechanism based on the CA attention mechanism to enhance feature extraction capability between the backbone network and the neck, strengthening the global and shallow features of the algorithm. This is beneficial for underwater unmanned platforms to extract small and difficult targets in water. Experiments show that the proposed network is feasible. Our algorithm reduces the number of parameters by 49.6% compared with YOLOX, and the accuracy on the URPC dataset is still improved, especially in the detection of small targets in the dataset. The results show that the algorithm has certain advantages and performs better in terms of parameter count and accuracy compared with the listed algorithms.

In our future work, we will further advance the balance between parameter count and accuracy in underwater object detection. We will also consider incorporating algorithm speed into the algorithm evaluation. The following methods will become important directions for our upcoming work: model pruning, numerical acceleration techniques, and loss function improvements.

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

**Funding:** This research was funded by National Natural Science Foundation of China, grant number 52171297.

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

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

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

**References**


33. Li, A.; Yu, L.; Tian, S. Underwater Biological Detection Based on YOLOv4 Combined with Channel Attention. J. Mar. Sci. Eng. 2022, 10, 469. [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.