ITC 2/52
Information Technology and Control
Vol. 52 / No. 2 / 2023
pp. 397-415
DOI 10.5755/j01.itc.52.2.33214

Underwater Object Detection Based on Improved Transformer and Attentional Supervised Fusion

Received 2023/01/15 | Accepted after revision 2023/03/18

HOW TO CITE: Li, Z., Li, C., Guan, T., Shang, S. (2023). Underwater Object Detection Based on Improved Transformer and Attentional Supervised Fusion. *Information Technology and Control*, 52(2), 397-415. https://doi.org/10.5755/j01.itc.52.2.33214

Underwater Object Detection Based on Improved Transformer and Attentional Supervised Fusion

Zhi Li, Chaofeng Li, Tuxin Guan

Institute of Logistics Science & Engineering; Shanghai Maritime University, Shanghai, 201306, China

Shaopeng Shang

Vocational College of Shanghai Jian Qiao University, 201306, Shanghai, China; e-mail: shaopengshang@126.com

Corresponding author: shaopengshang@126.com

Underwater object detection is one of the important technologies for improving the efficiency of underwater inspection, but the existing methods still suffer from the problems of missed detection and insufficient target localization capability of targets. To address these problems, an improved Transformer and multi-scale attentional supervised feature fusion-based underwater object detection method is proposed. In our method, the underwater objects are preprocessed by prior knowledge first. Then, a new coordinate decomposition window-based (CDW) Transformer block is proposed to extract spatial location information more accurately, and scaling factors are introduced to reduce the intermediate computation. Finally, an attentional supervised fusion (ASF) method is proposed to strengthen the link between feature extraction and feature fusion, and further improve the detected performance by using compound attention weights. The cascade detection head is improved, where the information flow is reversed to enhance the prediction of coordinates. The average accuracy of the proposed method on the URPC and DUO datasets is 3.7% and 3.8% higher than that of the baseline network through the cross-test, and outperforms the state-of-the-art methods. This study can provide a reference for engineering applications such as automated marine operations and biodetected fishing techniques.

KEYWORDS: Underwater Images; Object Detection; Transformer; Feature Fusion; Attention Mechanism.

1. Introduction

The ocean is rich in resources and has many unknowns; however, the research on marine is much more difficult than that on land. With the continuous exploitation of marine resources, the demand for automated underwater operations is increasing, which has caused underwater object detection technology to receive people's attention. Underwater object detection can be applied to aquaculture, fishing, environmental monitoring, underwater rescue and many other underwater tasks. Through this technology, advanced machines can realize autonomous underwater operations and greatly increase the efficiency of underwater operations.

In recent years, object detection algorithms have evolved from traditional manual feature extraction methods to deep learning methods. Traditional methods are highly affected by object and image acquisition, while deep learning methods use automatically extracted features to detect objects and can identify the physical characteristics of the objects more effectively [7].

Furthermore, the Transformer-based approaches have shown better performances in object detection [8]. Transformers consist of self-attention and allow for better modelling of the relationships between all pixels [4]. Increasingly, researchers use Transformers to obtain better models than just convolutional neural networks (CNNs), which gives us the motivation to introduce the Transformer for underwater object detection.

The existing deep learning-based methods have achieved good performance in common scenes. However, when encountering complex underwater environments, these methods still exhibit missing detection and insufficient positioning ability for hard objects. Different from ordinary detection, the difficulty of underwater object detection comes from various disturbances and motion blur during visual data acquisition, which causes the existing problems. For the disturbances, some work reduced the noise through image enhancement, and others improved detection accuracy by adjusting the network structure. However, the motion blur is easily overlooked. Existing work often treats these two different sources of difficulty together, which leads to poor model generalization performance. In our work, we take motion blur as prior knowledge, and learn the disturbances in underwater images through an improved network.

The primary goal of this research is to reduce missed detection in underwater scenes and enhance the precise positioning capabilities, introducing Transformer to improve detected accuracy. The main contributions of this paper are as follows:

1. A coordinate decomposition window (CDW) Transformer block is proposed for the precise object positioning, which strengthens the spatial position information by the novel coordinate decomposition calculations and reduces the intermediate computation by the scaling factors.
2. An attentional supervised fusion method (ASF) is proposed for the Transformer and incorporated into the feature pyramid to enhance small objects features, which can supervise multi-scale information fusion and strengthen the link between feature extraction and feature fusion.
3. During training, prior knowledge is used for data preprocessing to alleviate underwater motion blur, and the information flow in the cascade detection head is reversed to enhance coordinate prediction.

The rest of this paper is arranged as follows. Section 2 presents some related works. Section 3 describes the proposed method. Section 4 presents the experimental results and analysis. Finally, the conclusion is presented in Section 5.

2. Related Work

2.1. Underwater Object Detection

Underwater object detection can be divided into two categories according to the different signals of the object The first category is the acoustic image collected by sonar [26]. The second category is optical images acquired by cameras. Sonar images usually have noisy data of lower resolution and more difficult interpretation, so our approach aims to underwater optical images.

Currently, underwater object detection algorithms can be roughly divided into two categories: handcrafted feature-based methods and deep learning-based

methods. As an example of the former type, Huang et al. [9] extracted a feature combination of colour, shape and texture properties to detect fishes. Weber's Local Descriptor (WLD) [29] and Histogram of Oriented Gradients (HOG) [11] descriptor were used early to assist detection. However, these methods need specific characteristics and are far from meeting the needs of practical applications.

Deep learning-based methods are widely used in various fields [5], including underwater detection. The SWIPENet model [2] utilized a sample reweighting scheme to reduce environmental interference, significantly improving the accuracy of small object detection. Dawid Połap et al. [26] analyzed the sonar images according to regions of interests (ROIs) by the histogram module to exclude non-target images and then classified the objects by CNN, which achieved good results in real scenarios. The RoIMix network [17] mixed the proposals extracted from different images to improve the detection of small underwater objects. This type of method has come to be further divided into one-stage and two-stage methods and has made some outstanding achievements, respectively. For example, the R-CNN series of networks have been applied to underwater object detection for a long time. Li et al. [12] first introduced Fast R-CNN to fish species detection. Mandal et al. [23] first introduced an end-to-end deep learning-based architecture that used Faster R-CNN to detect fish. However, these methods are less effective and slower in detecting when the targets overlap. In one-stage detection methods, an improved YOLOv3 network was used with data augmentation methods to increase the domain diversity of the dataset and the speed of detection [21]. Zhang et al. [34] focused on lightweight performance and proposed a lightweight underwater object detection method based on the YOLOv4. Lei et al. [10] first used YOLOv5, Swin Transformer and adjusted the feature fusion to achieve high detection accuracy. Recently, Yan et al. [33] improved the YOLOv7 network, incorporating the CBAM attention mechanism to enhance features and achieving optimal results for one-stage underwater detection methods. However, the method is not accurate enough for detecting tiny targets. Overall, the attention mechanism is popular for enhancing the ability of underwater target feature extraction, but there is still much room for accuracy improvement. We proposed a novel Transformer block and an attentional supervised fusion method to handle missing detection.

2.2. Combination of Transformer and CNN Techniques

In recent years, the combination of Transformers with convolutional neural networks (CNNs) have been shown better detection results [15]. The convolutional operation in CNNs tends to focus on local feature processing, but the receptive field is usually small. In contrast, Transformers obtain the global receptive field by computing the global correlation, but the local details are weaker. To overcome these challenges, an approach based on the fusion of these two aspects has been widely studied.

Broadly speaking, the DETR network and its variants [1] combine the CNN backbone with the Transformer encoder and decoder to interpret the relationship between the object and the global image, omitting postprocessing and outputting the final result directly. The authors of [20] introduced a hierarchical structure commonly used in CNNs to build a multilayer Transformer structure to enhance multiscale detection. Narrowly speaking, the combination of Transformer and CNN methods can be attributed to the combination of self-attention and convolution. For example, UniFormer [14] alternately uses convolution and self-attention computations, while the ACmix model [24] reunifies the mapping part to combine convolution and self-attention mechanisms more effectively. These works show that the combination of the Transformer and CNN methods can enable the network to inherit the advantages of both global and local features. Inspired by these works, the proposed CDW Transformer block explores the advantages of global and local information, improving the detection performance in complex underwater scenes. Unlike other works, CDW is more capable of processing spatial information and more flexible in intermediate calculations.

2.3. Multi-scale Feature Fusion

Multiscale feature fusion has been widely used in object detection. For example, Lin et al. first proposed a feature pyramid network (FPN), which uses top-down channels to build high-level feature maps at all scales [16]. PANet [18] added a path to obtain more details of objects. Aug-FPN [6] adaptively reduced the

loss of contextual information in high-level feature maps and accelerated inference. Bi-FPN [27] added various channels to the FPN to fuse features at different stages by different weights. These methods add various auxiliary means in the feature fusion stage to achieve a better fusion effect, but too many connections make the model computationally intensive.

However, these feature fusion processes are designed for CNN's feature map, and the effect on the Transformer's attention map of features is not as good as that in CNN. Therefore, a better method for adapting to the Transformer's attention fusion needs to be explored. We propose a new attention supervision fusion method, which designs a path supervised by the attention weight of the Transformer block on the basis of the FPN. The proposed ASF not only improves the feature extraction but also strengthens the link between feature extraction and feature fusion.

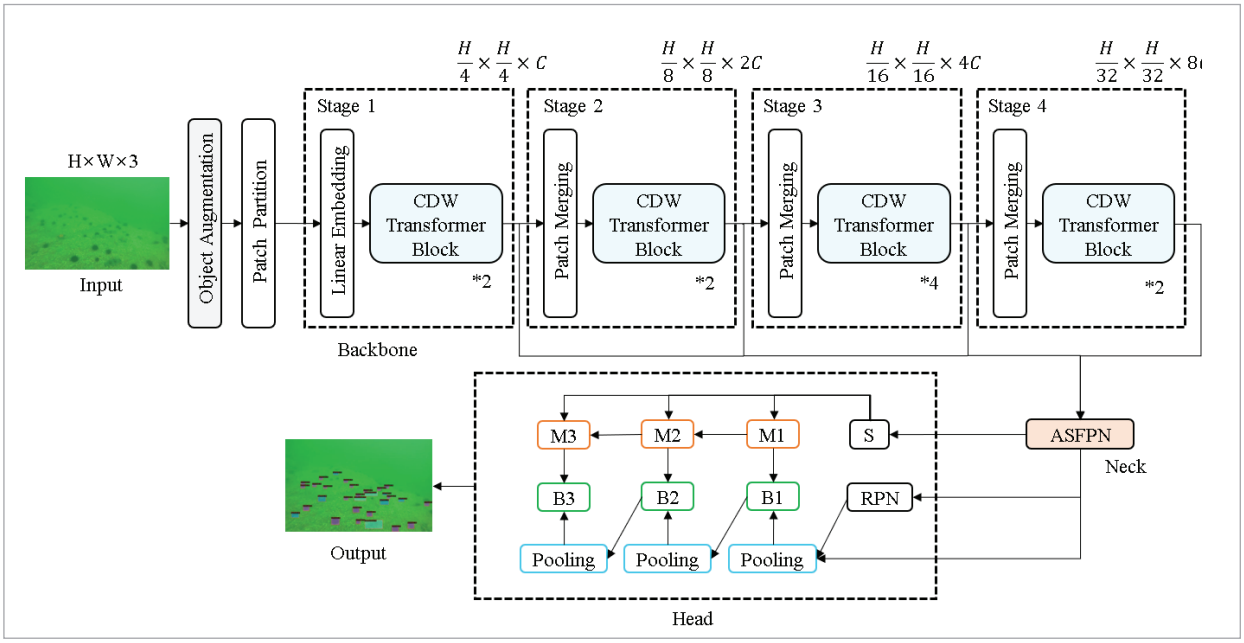
3. Proposed Method

The difficulty in underwater object detection is more about missed detection rather than false detection, so the model needs to focus on the spatial information and object features. The overall architecture of our coordinate decomposition window-based Transformer network (CDWTN) is shown in Figure 1. First, the underwater image data are preprocessed for object augmentation. Then, the feature is extracted by the improved Transformer blocks and processed by the attentional supervised fusion pyramid network (ASFPN). Finally, the detection results are output by the cascaded detection head.

3.1. Improved Transformer Feature Extraction Network

As shown in Figure 1, to maintain the detection capability for large underwater images, the backbone of the proposed method follows the design of four cascaded stages [30]. The model initially starts by slicing the image into nonoverlapping image patches, and then maps it into windows of arbitrary dimension using a patch partition layer and a linear embedding layer. The feature map size in the first stage is H/4×W/4, and the number of output channels is C. In the following three stages, each stage consists of multiple CDW Transformer blocks and a patch merging [20] layer to reduce the size of the feature maps. The downsampling in the hierarchical Transformer allows the

Figure 1
The overall architecture of the CDWTN



model to gradually extract global information and has better scale adaptability.

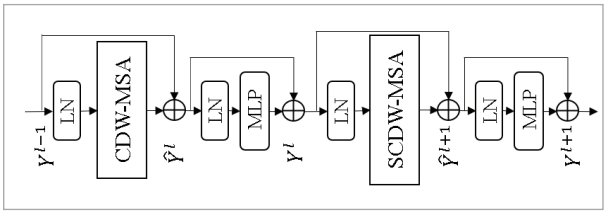
3.2. CDW Transformer Block

The proposed transformer block uses the decomposition coordinate information to strengthen the spatial position information in the self-attention, and effectively reduces the intermediate parameters by the scaling factor.

The structure of the proposed CDW Transformer block is shown in Figure 2. The feature map after layer normalization (LN) is fed into multi-head self-attention blocks for feature learning and then fused with the residual-connected feature map. The SCDW-MSA block adds a shift operation [20] to the CDW-MSA block. The multi-head approach superposes single self-attention results in parallel. The entire feature extraction network is constructed with alternating window-based and shifted window-based CDW operations interacting between image patches within each stage.

Figure 2

CDW Transformer block, where CDW-MSA denotes coordinate decomposition window multi-head self-attention and SCDW-MSA denotes shifted coordinate decomposition window multi-head self-attention



Specifically, to emphasize spatial information, CDW integrates self-attention and convolution operations by introducing decomposed coordinate information into the self-attention computation. This integration is applicable to images of various sizes and reduces information loss due to image chunking, thus allowing the backbone network to better generate attention weights. The continuous Transformer blocks are computed as Equations (1)-(4):

$$\hat{Y}^{l} = \text{CDW-MSA}(\text{LN}(Y^{(l-1)})) + Y^{(l-1)} \quad (1)$$

$$Y^{l} = \text{MLP}(\text{LN}(\hat{Y}^{l})) + \hat{Y}^{l} \quad (2)$$

$$\hat{Y}^{(l+1)} = \text{SCDW-MSA}(\text{LN}(Y^{l})) + Y^{l} \quad (3)$$

$$Y^{(l+1)} = \text{MLP}(\text{LN}(\hat{Y}^{(l+1)})) + \hat{Y}^{(l+1)} \quad (4)$$

The $\hat{Y}^{l}$ and Y^{l} represent the output of the CDW Transformer block and the MLP block, respectively. The specific improvements are introduced as follows.

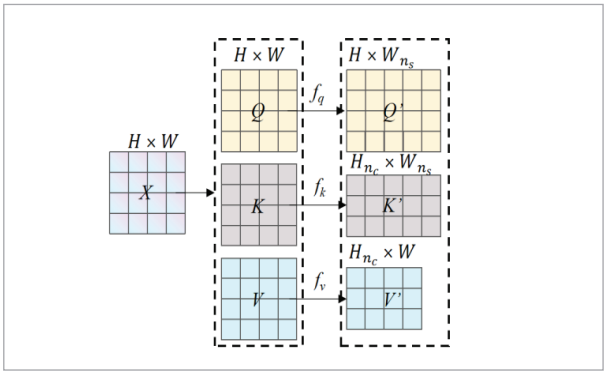
3.2.1. Scaling Factor

Since the traditional self-attention module calculates all pixel points and the object is widely distributed in the underwater environment, there is some redundant information in the calculation. Therefore, to lighten the model, two scaling factors n_c and n_s are introduced into the Transformer blocks.

The query matrix (Q), key matrix (K), and value matrix (V), which have identical sizes after mapping the input image of height H and width W, are transformed by functions $f_q(x)$, $f_k(x)$ and $f_v(x)$. The transformed matrices are the adjustable matrices *Q'*, *K'* and *V'* shown in Figure 3.

Figure 3

Diagram of the scaling factor



In regard to the scaling factors, it is experimentally demonstrated that the adjusted image matrix does not degrade the detection accuracy when reducing computational overhead.

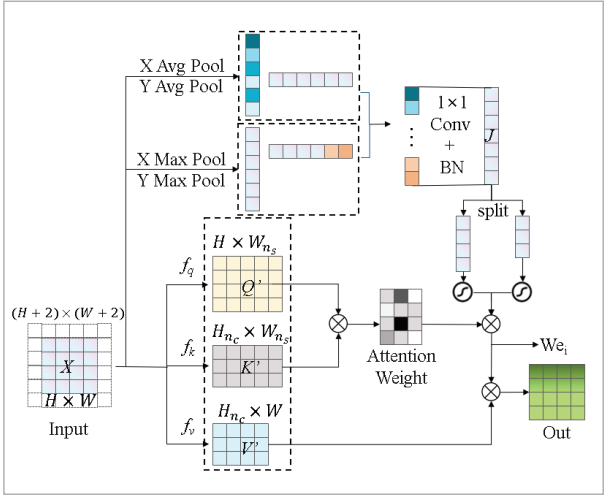
3.2.2. Coordinate Decomposition Window-based Self-attention

In some Transformer structures, the image is sliced and shifted to avoid excessive computational complexity. However, in practical underwater scene applications, image slicing will result in poor detection

accuracy for objects that are located in boundary regions and patch intersections. To address this problem, the method presented in this paper uses coordinate information to enrich the expressiveness of features. The method strengthens the weights of object locations and reorganizes self-attention and convolution. The pooling operation is used to hasten the receptive fields to enable an association with the surrounding image patches. The single-channel coordinate decomposition window-based self-attention calculation is shown in Figure 4. The upper half is the coordinate decomposition branch, and the lower half is the self-attention branch. While calculating the self-attention, the horizontal and vertical coordinate decomposition information of the input and the neighbourhood, are calculated in the $(H+2)\times(W+2)$ region, respectively.

Figure 4

Coordinate decomposition window-based self-attention



Specifically, the coordinate decomposition branch is used to obtain the four decomposition feature codes in the horizontal and vertical directions by averaging pooling and maximum pooling. Such sampling methods optimize the feature decomposition richness. After decomposition pooling, the output of the c^{th} channel at height h and the output of the c^{th} channel with width w can be expressed as Equations (5) and (6), respectively. Each element in the decomposition feature encoding reflects whether the exists in the corresponding row and column. Next, the four feature encodings are transposed and spliced by the concat approach. Then the pixels are reduced using a shared 1×1 convolution and normalized with batch normalization (BN) [28] to generate intermediate attention weight J. Finally, weight J is split into two separate matrices J^w and J^h along the spatial dimension, and the sigmoid mapping function is used to add nonlinear properties, which smooths the gradient and avoids jumping output values. The classical sigmoid maps any real-valued number to a value between 0 and 1, which can be interpreted as a probability. In the CDW Transformer block, the sigmoid function outputs the probability that the coordinate location has an object.

$$Z_c^w(w) = \frac{1}{H} \sum_{0 \le j < W+2} x_c(j, w) \tag{5}$$

$$Z_c^h(h) = \frac{1}{W} \sum_{0 \le i < W+2} x_c(h, i) \tag{6}$$

The self-attention branch is calculated using the deformed matrix. The attention weights are then multiplied by the coordinate decomposition weights to obtain the composite attention weight W_{ei}. W_{ei} is used for the feature fusion in the neck and will be weighted to the V' matrix.

Through a richer and more flexible attention calculation, the CDW Transformer block extracts richer object information during the feature extraction process. The coordinate decomposition calculation we designed can also be understood as a positional attention mechanism, which is used to strengthen the target position information. It not only preserves the details of the underwater objects through self-similarity but also enhances the location information of the object.

After combining the coordinate decomposition calculation and the original self-attention, Equation (7) shows the calculation in the CDW Transformer Block.

$$\text{CDW-SA}(Q', K', V', J') = \text{softmax}\left(\frac{Q'(K')^T}{\sqrt{d_k}}\right) V' J^w J^h \tag{7}$$

The Q', K' and V' are the adjustable matrices. J^w and J^h are the weight splits of J. As with all self-attention mechanisms, the softmax maps the output of multiple neurons into the (0,1) interval. The d_k is the embedding dimension to represent each entity, which is used to prevent large-scale inputs from excessively affecting the calculation of weights.

CDW-MSA is an extension of CDW-SA in which we run k self-attention operations in parallel. In Equation (8), U_{msa} represents an output matrix related to the size of the intermediate mapping matrix [1].

$$\text{CDW-MSA}(Q', K', V', J') = [\text{CDW-SA}_1(z); \text{CDW-SA}_2(z); \cdots; \text{CDW-SA}_k(z)]U_{msa} \quad (8)$$

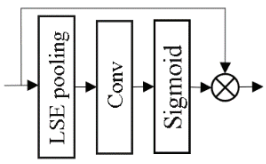
3.3. Attentional Supervised Fusion and Cascaded Detection Head

3.3.1. Attentional Supervised Fusion

This paper designs an ASF to integrate multi-scale features in the neck stage. The closest work to our method is PANet, but the connection in PANet is unstable. Therefore, we design a new supervision method that connects the feature extraction and fusion of transformers. In addition, the coordinate decomposition information in CDW is retained to supervise the feature fusion of the object region.

Inspired by spatial attention [32], the single-scale attention-supervised feature fusion introduces the novel spatial selection attention (SSA) module, as shown in Figure 5. The optimal combination of the average and maximum pooling results is obtained adaptively by log-sum-exp (LSE) pooling [25]. The weight W_{ei} is filtered by LSE pooling, residual concatenation, sigmoid mapping function and 7×7 convolution, which is consistent with the backbone window size, so that pixels with similar scores obtain similar weights in the training process.

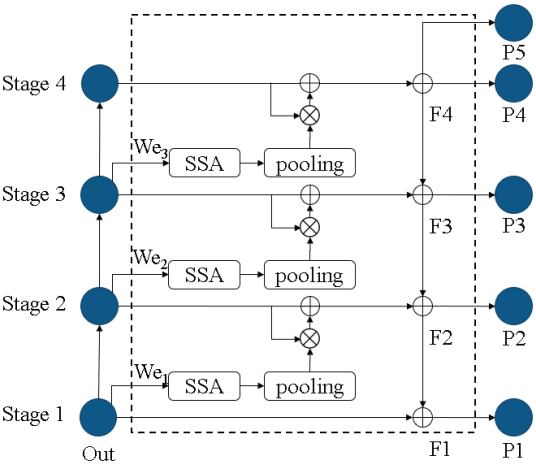
Figure 5
The structure of SSA



Unlike most existing feature fusion structures, we extract the composite attention weights in the backbone for supervising the guided fusion process, which provides a certain regularization effect for the network. The parameters used here other than the SSA module can be shared with the backbone, so the burden on the odel is small.

When further extended to multi-scale features, this supervision method is incorporated into the feature pyramid as shown in Figure 6.

Figure 6
The structure of ASFPN



First, the output feature maps and the composite attention weights of each stage of the improved Transformer feature extraction network are input into the ASFPN. The weight parameters are then filtered and used to enhance the original output feature maps. This approach reinforces the connection between the feature extraction and feature fusion processes. The filtered weights are pooled and then multiplied and summed with the output feature maps of the final layer. Intermediate feature maps [F1, F2, F3, F4] are obtained. Then, four sets of feature maps [P1, P2, P3, P4] are obtained from the reverse path, and the resolution is reduced again to obtain P5. Finally, five sets of feature maps are sent to the detection head for category and location prediction. In other words, the fusion process is supervised and enhanced by adding attentional supervision that uses filtered weights to the feature pyramid. The attention supervision pathway can enhance the multi-scale features of the target area in a targeted manner. The SSA module focuses on the spatial information related to the object in the attention weights to reduce the loss of small underwater objects in the resolution reduction process.

3.3.2. Cascaded Detection Head

The final detection results are obtained by means of the adjusted cascade detection head. The detection head component refers to the HTC framework [3], which demonstrated the feasibility of reinforcing the information flow between mask branches by feeding the mask features and the box features of the preceding stage. Similarly, we adjusted the direction of the information flow, using the simplified mask information flow as the inputs of the box regression branch to form a coordinate information flow. The cascaded detection heads are connected as shown in the head section on the lower side of Figure 1. The coordinate information flow branch includes semantic head S [3] and coordinate information M_i. The lower half still uses a region proposal network (RPN), pooling and bounding box regression B_i. If the coordinate information flow exceeds the bounding box, the prediction bounding box will be fine-tuned, while the information flow loss is added to the overall loss. This improves underwater object detection by combining the advantages of the cascade approach and the complementary nature between object frame and mask coordinate prediction. Other structures are consistent with HTC. We apply SoftNMS to the box results, which makes the model more efficient. In each head block, the box head predicts the classification score c_i and regression offset r_i. The cross-entropy (CE) loss and smooth L1 loss are used to calculate box loss (L^i_b) and semantic loss (L_s). Smooth L1 loss improves the generalization of the model. The information flow loss (L^i_m) is calculated by the binary cross-entropy (BCE) loss. We set α = [1, 0.5, 0.25] and β = 1. The overall loss function takes the form of multi-task learning [3]:

$$L = \sum_{i=1}^{T} \alpha_i (L_b^i + L_m^i) + \beta L_s \tag{9}$$

$$\begin{aligned} L_b^i(c_i, r_i, \hat{c}_i, \hat{r}_i) &= L_{cls}(c_i, \hat{c}_i) + L_{reg}(r_i, \hat{r}) \\ &= CE(c_i, \hat{c}_i) + smooth_{L1}(r_i, \hat{r}) \end{aligned} \tag{10}$$

$$L_m^i(m_i, \hat{m}_i) = BCE(m_i, \hat{m}_i) \tag{11}$$

$$L_s = CE(s, \hat{s}) \tag{12}$$

4. Experiment and Analysis

4.1. Experimental Data

4.1.1. Underwater Dataset

We use the 2020 Underwater Robot Professional Contest (URPC) and DUO [22] object detection datasets for testing. In URPC, there are 5543 images covering five categories of objects: holothurian, echinus, scallop, starfish and water plants. After data screening according to official requirements, the final dataset contained 5455 images without water plants. Since no additional test sets were announced, while ensuring no overlapping data, the dataset was divided randomly at a ratio of 9:1, i.e., 4909 images in the training set and 546 images in the test set. In the DUO dataset, there are 7782 images covering four categories: sea urchins, sea cucumbers, scallops, and sea stars, with 6671 fixed images in the training set and 1111 fixed images in the test set.

The download addresses of the datasets used in this article are as follows:

URPC2020 https://openi.pcl.ac.cn/OpenOrcinus_orca/ URPC2020_dataset;

DUO https://drive.google.com/file/d/1w-bWevH7jFs7A1bIBlAOvXOxe2OFSHHs/view?usp=sharing.

4.1.2. Data Processing

Image quality has a profound impact on subsequent work [13]. Different from other environments, the underwater images have many problems such as noise interference, blurred texture features, low contrast and colour distortion. Therefore, the underwater target detection task faces many challenges.

Existing underwater image enhancement methods focus on the various noises in the image, ignoring the motion blur caused by the image acquisition process and water movement. In this paper, the blurring problem of images is considered, which complements existing methods.

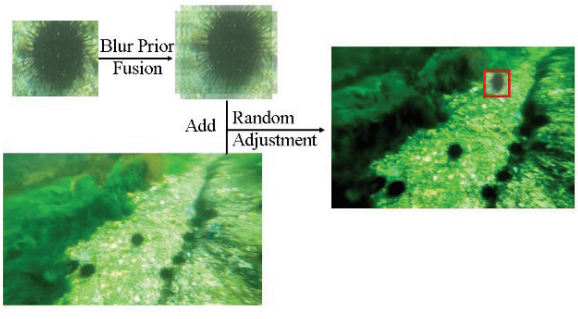
For motion blur and detailed loss in underwater images, we introduce prior knowledge and the mosaic method to augment the objects. We introduce the method using local minimal pixel prior [31] to quickly deblur the image. The detected effect before and after deblurring is shown in Figure 7, which suggests that the recovered images are effective in improving the detected accuracy of some scenes.

Figure 7
Detected results before and after deblurring

(a) before deblurring (b) after deblurring

For some blurs that cannot be quickly removed, we design a blur prior fusion method to adapt to these blurs simply and effectively. Objects in the dataset are selected for pixel panning and proportional overlay to create a motion blur. Pixels are randomly shifted at a distance of no more than 1/20 of the object size; otherwise, ghosting and duplicate recognition will occur. The background is cropped, and the brightness and three-channel histogram of the images are randomly adjusted. An example of blur prior fusion is shown in Figure 8.

Figure 8
Example of blur prior fusion



4.2. Model Training and Evaluation Metrics

In this paper, the BN layer in the detection head is replaced with a group normalization (GN) layer to accelerate the convergence and reduce the accuracy reduction when the batch size is small. The number of the CDW Transformer blocks in each stage is set to 2, 2, 4, and 2, and the ImageNet-1K pretrained model is used to accelerate training. The window size is set to 7 by default. The Adam optimizer (initial learning rate of 0.0001 and weight decay of 0.05) is used to optimize the training process. Under the same experimental conditions, the input image size of the models was set to 512 × 512, while the convergence was accelerated using the respective pretrained models. The batch size is set to 4. The training and testing process was performed on an NVIDIA GeForce RTX 3090 GPU with 12 GB of RAM.

In this paper, the average precision accuracy (AP) and the mean average precision (mAP) are calculated as evaluation metrics. The mAP includes AP@0.5 and AP@0.5:0.95. AP@0.5 indicates the average accuracy when the threshold value of the intersection and merge ratio Intersection over Union (IoU) between the detection area and the object area is 0.5. AP@0.5:0.95 indicates that when IoU increases from 0.5 to 0.95, the average AP is calculated with a step size of 0.05. For AP@0.5:0.95. The higher the IoU threshold is, the greater the coincidence ratio is between the result box and the calibration box. This index requires the model to have the better object location ability. The mAP of m types of objects is calculated by Equation (13):

$$mAP = \frac{1}{m}\sum_{i=1}^{m} AP_i \tag{13}$$

4.3. Comparative Experiments

The proposed model is tested on the open URPC2020 dataset and DUO dataset. Six pure CNN methods and five methods containing the Transformer are selected for comparison.

The results are shown in Table 1. We perform 10 cross-validations on the URPC dataset, and the results are given in Table 2. From Table 2, it can be seen that the accuracy change is only 0.4%, which suggests the stability of the model.

It can be seen in the experimental results in Table 1 that the AP@0.5 index of CDWTN is 88.3% and 92.8%, which is better than that of the other models. Compared with the optimal network of CNN on the URPC data, the AP@0.5 index is 1.4% higher than that of the state-of-the-art architecture, i.e., YOLOv7. The AP@0.50:0.95 index is 5.1% higher than that of the YOLOv7 network, and the model object positioning is significantly more robust and meets the detection requirements for practical applications. Compared with Transformer class networks, the proposed method

Table 1
Performance comparison on the URPC2020 and DUO datasets

Categories	Method	Backbone	URPC		DUO	
			AP@0.5 %	AP@0.50: 0.95 %	AP@0.5 %	AP@0.50: 0.95 %
CNN	Faster R-CNN	ResNet-50	65.2	35.0	67.7	37.1
	RetinaNet	ResNet-50	68.7	43.4	69.9	45.0
	YOLOv3	DarkNet53	78.3	46.2	79.8	47.6
	YOLOx	CSPDarkNet	80.2	48.0	85.0	52.4
	YOLOv5-s [10]	CSPDarkNet53	80.1	45.6	86.6	57.9
	YOLOv5-l [10]	CSPDarkNet53	85.1	49.8	87.9	58.5
	YOLOv7 [33]	CSPDarkNet53	86.9	51.7	89.5	60.6
Transformer	DETR [1]	ResNet-50	65.5	38.7	76.7	50.9
	Deformable Detr [35]	ResNeXt-101	74.4	45.0	80.1	56.1
	DAB DETR [19]	ResNeXt-101	79.6	48.9	88.1	61.7
	HTC	ACmix [24]	85.8	53.3	88.4	62.8
	HTC	SwinT [20]	85.6	53.3	87.9	62.1
	Ours	CDW	**88.3**	**56.8**	**92.8**	**67.3**

Table 2
Cross-validation on URPC2020

Group	1	2	3	4	5	6	7	8	9	10
AP@0.5	56.8	56.4	56.6	56.6	56.7	56.8	56.4	56.8	56.8	56.5
AP@0.50:0.95	88.3	88.1	88.3	88.3	88.2	88.4	88.2	88.3	88.3	88.2

has a 2.5% higher AP@0.5 index than the novel ACmix backbone. The Transformer feature extraction method is adopted to enhance the robustness of different backgrounds through self-attention calculation, which is also more suitable for the complex underwater environment. Moreover, CDW blocks are adopted to improve the richness of feature extraction, and the convolution is integrated to focus on the location information of the object. The scaling factor is introduced in addition to adding coordinate information, which makes the calculation of self-attention simpler and faster. Finally, ASFPN makes the fusion process subject to the supervision of Transformer backbones. Weight parameters are used to supervise the location and characteristics of objects in the fusion process, thus reducing the attention to other interference information, making the information of small objects in deeper networks better fused, and improving the effectiveness of multi-scale object detection.

Furthermore, we perform the cross-test using the training and test sets of the two datasets to verify the generalization performance of the model, the results of which are shown in Table 3.

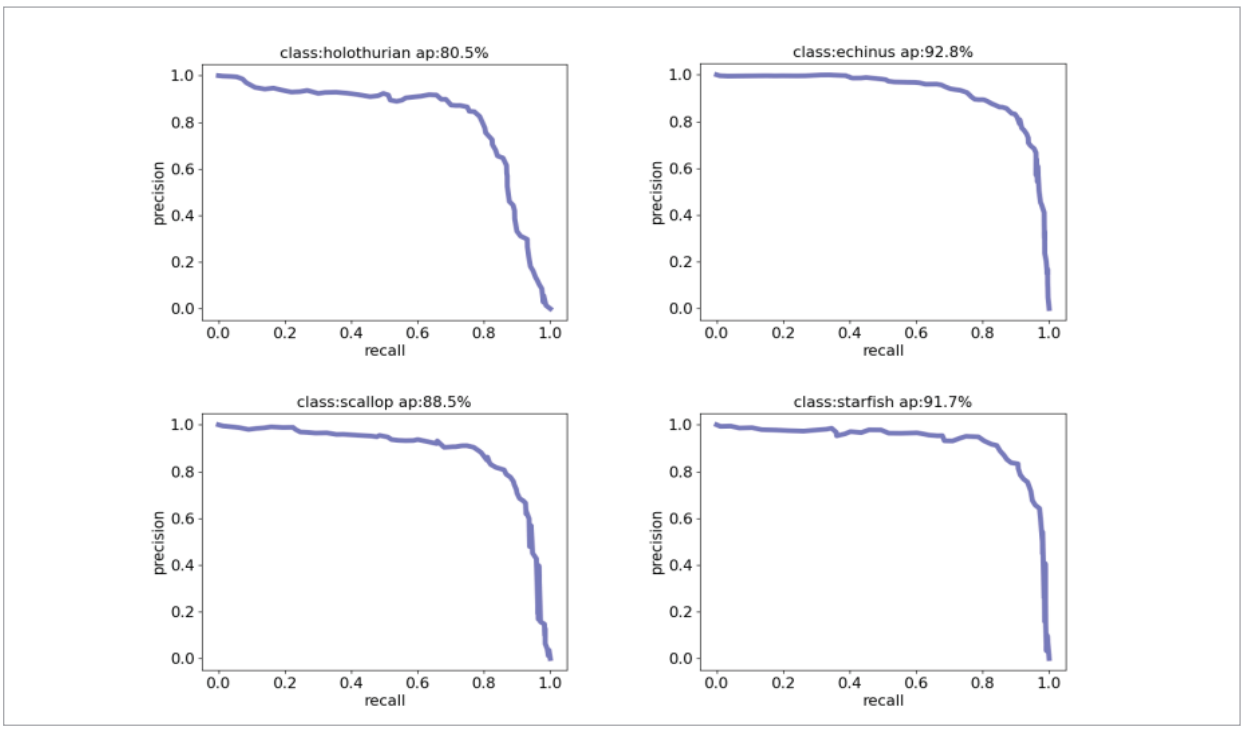
From Table 3, the test accuracy has decreased due to the data for different distributions, but the accuracy is still higher than other methods. The CDWTN inherits the flexibility of hierarchical transformers to model at different scales and has flexible intermediate calculations. In addition, blur prior fusion also improves generalization performance. Furthermore, we plotted the PR curves of various objects.

Figure 9 shows the PR curves of the proposed CDWTN for each class of underwater objects, where

Table 3
Performance comparison of different methods via cross-validation

Categories	Method	Backbone	Train/test (URPC/DUO)		Train/test (DUO/URPC	
			AP@0.5 %	AP@0.50: 0.95 %	AP@0.5 %	AP@0.50: 0.95 %
CNN	Faster R-CNN	ResNet-50	55.6	30.2	60.3	35.3
	RetinaNet	ResNet-50	60.3	37.1	66.5	43.5
	YOLOv3	DarkNet53	68.9	40.5	77.1	46.4
	YOLOx	CSPDarkNet	70.1	43.0	78.6	47.1
	YOLOv5-s [10]	CSPDarkNet53	73.3	42.9	78.2	45.4
	YOLOv5-l [10]	CSPDarkNet53	79.1	46.3	82.1	49.7
	YOLOv7 [33]	CSPDarkNet53	81.4	48.6	84.8	51.8
Transformer	DETR [1]	ResNet-50	61.3	35.1	66.7	39.8
	Deformable Detr [35]	ResNeXt-101	72.5	42.2	75.3	46.1
	DAB DETR [19]	ResNeXt-101	76.8	45.8	80.5	49.8
	HTC	ACmix [24]	82.3	50.4	85.8	53.4
	HTC	SwinT [20]	81.6	50.4	85.7	53.3
	Ours	CDW	**85.4**	**53.1**	**89.4**	**57.3**

Figure 9
PR curves of various objects



P represents the precision and R represents the recall. The CDWTN has the best detection ability for echinus, followed by starfish. The accuracy of the optimal model for both types of networks is also given for each category, as shown in Table 4. The accuracy of the proposed network is improved for each category, especially for the scallop category, which is 3.1% higher than that of the YOLOv7 network and 6.7% higher than that of the SwinT method.

Table 4
Results for various classes from the URPC2020 test set

Method	AP			
	Holothurian	Echinus	Scallop	Starfish
YOLOv5-l	77.6	91.6	83.0	88.2
YOLOv7	80.5	91.4	85.4	90.3
SwinT	78.9	91.2	81.8	90.5
CDWTN	**80.5**	**92.8**	**88.5**	**91.7**

4.4. Ablation Study

In this paper, to verify the module effectiveness, each modified version of the CDWTN underwater object detection method is tested separately on the URPC2020 dataset, where Baseline denotes the SwinT method.

Baseline + A-data denotes the use of the pro-processed methods. Baseline + nc&ns denotes the introduction of scale factors. Baseline + CDW denotes the use of the improved coordinate decomposition self-attention calculation. Baseline + ASFPN denotes the use of the ASF method. CDWTN denotes the proposed final method. The results of the ablation experiments are shown in Table 5, where AP_s are set as small object AP values with pixels smaller than 32^2, AP_m refers to medium object AP values with object pixels between 32^2 and 96^2, and, AP_l refers to large object AP values with pixels larger than 96^2.

Experiments of A-data. As shown in Table 4, Baseline + A-data enriches multiscale objects by prior knowledge. The indexes AP@0.5 and AP@0.5:0.95 increase by 0.5% and 1.1%, respectively, compared with the baseline network. It can be found that the increase in object volume is beneficial to network learning. The object is richer in the local scope of the picture by using blur prior fusion.

Experiments of CDW Transformer blocks. As shown in Table 5, with the introduction of the scale factors, the number of intermediate calculations can be reduced and has little influence on accuracy. Baseline + CDW introduces a scaling factor and coordinate decomposition window-based self-attention, which improve objects of all scales to a certain extent. The introduction of coordinate decomposition attention can better extract the features of objects in fuzzy images with rich image backgrounds and can strengthen the ability to extract objects.

Table 6 shows the experimental results with different quantities of Transformer blocks, with the number of

Table 5
Ablation experimental results

method	AP@0.5	AP@0.50:0.95	AP_s	AP_m	AP_l
Baseline	85.6	53.3	25.5	46.3	58.0
Baseline + A-data	86.1	54.4	26.8	47.4	58.6
Baseline + n_c& n_s	85.6	53.2	25.4	46.2	58.1
Baseline + CDW	86.0	54.0	26.3	47.0	58.3
Baseline + ASFPN	86.4	54.6	27.9	48.3	58.8
Baseline + A-data + CDW	87.5	55.2	27.8	48.7	59.5
Baseline + A-data + ASFPN	87.3	55.8	29.2	50.5	59.2
Baseline + CDW + ASFPN	86.9	55.6	29.0	50.8	59.1
CDWTN	**88.3**	**56.8**	**29.7**	**51.1**	**60.1**

Table 6
Experimental results with different quantities of CDW Transformer blocks

Number of Transformer blocks in four stages	AP@0.5	AP@0.50:0.95	FPS f/s
2, 2, 2, 2	80.6	49.3	**13.4**
2, 2, 4, 2	86.0	54.2	11.3
2, 2, 6, 2	**86.6**	**54.7**	8.0

frames per second (FPS) transmitted as a reference metric. The number of blocks in each phase should be an even number of blocks, following the principle of alternate use. As shown in Table 6, three block number ratios are demonstrated, and a better combination of speed and accuracy can be achieved by reducing the number of blocks in the third stage to 4 within the range of accuracy allowed. After adding the deformation factor, the accuracy of the model remains good, while offsetting the computational complexity caused by other additional parameters.

Experiments of ASFPN. The introduction of the ASFPN module Baseline + ASFPN reduces the loss of small objects caused by feature fusion. Compared with the baseline network, AP@0.5 and AP@0.5:0.95 increase by 0.8% and 1.3%, respectively, and the detection accuracy of small objects AP_s increases by 2.4%, indicating that replacing the original FPN feature fusion method with ASFPN can enhance the small object detection capability. ASFPN uses a supervised approach that utilizes attention to focus on the location of objects, making the feature extraction and feature fusion fitter.

To further demonstrate the feature fusion effect, this paper compares the same type of feature fusion methods on the baseline network, as shown in Table 7.

Table 7
Performance comparison of different feature fusion methods

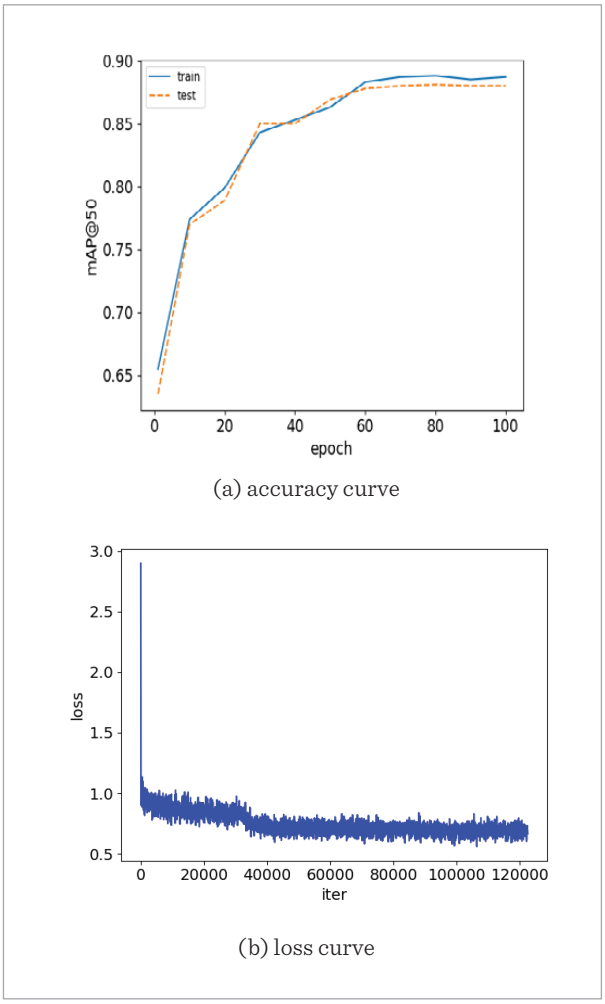
Feature Fusion Method	AP@0.5	AP@0.50:0.95	AP_s
FPN [16]	85.6	53.3	25.5
PANet [18]	85.8	53.8	26.7
Bi-FPN [27]	86.2	54.1	26.3
ASFPN	**86.4**	**55.6**	**27.9**

Among them, PANet and Bi-FPN can bring great improvement in convolutional networks. However, the improvement brought by ASFPN for the Transformer network is more effective for improving underwater detection accuracy. Finally, the best detection performance is obtained by integrating all the above improvements of the proposed method.

4.5. Visualization Results

We plotted the accuracy curve (mAP@50) for every 10 epochs and the loss curve for every 25 iterations. In Figure 10, it can be seen that the model is constantly converging, and is not overfitting. Since the batch size setting is small, the glitch of the loss curve is noticeable.

Figure 10
Accuracy curve and loss curve



(a) accuracy curve

(b) loss curve

In addition to the quantitative comparison, some visualization results are shown in Figures 11 and 12. To facilitate the display, the detection results are marked using transparent rectangular blocks. Among them, holothurian, echinus, scallop and starfish are shown with dark blue, purple, brown and light blue transparent rectangular blocks, respectively, with the aid of white boxes. The CDWTN network proposed in this paper, to a certain extent, solves the problem of missed detection and false detection of image edge objects existing in the baseline network and makes a great improvement in the detection of small objects and fuzzy objects. However, there is still room for improvement for some small objects that cannot be discerned by the naked eye. Figure 11 shows the good detection results in some blurred scenes.

Figure 12 shows a partial comparison of the detection results between the method presented in this paper

Figure 11
Detected results of fuzzy scene

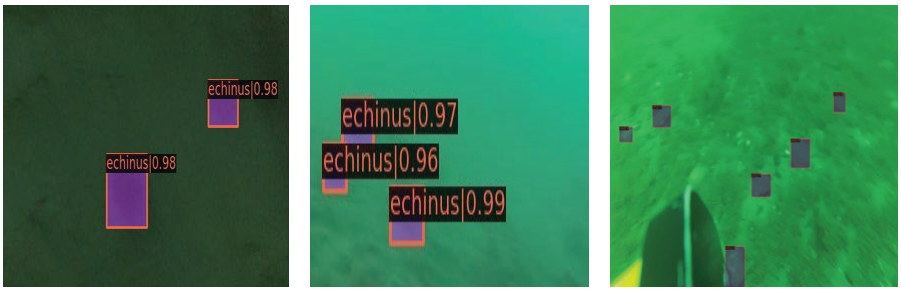
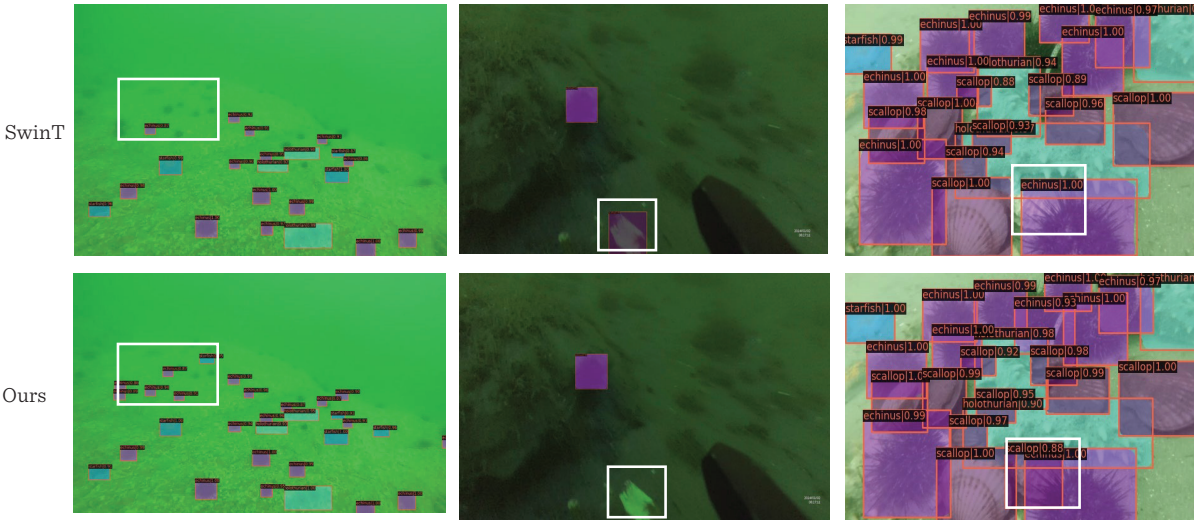


Figure 12
Comparison of the detection results for SwinT and CDWTN



(a) Small Objects (b) Low Light (c) Overlapping Objects

and the SwinT method. Figure 12(a) demonstrates that the method makes a significant improvement in the detection accuracy for small objects in blurred scenes. The proposed method improves the information interaction between image patches. It also benefits from the attentional supervised feature fusion process, which allows small object features to be retained in deeper networks when resolution is reduced. Figure 12(b) shows a low-light image with a large original image size. The proposed method reduces the false detection of scallops with the presence of varying degrees of sediment obscuration. The improved method increases the learnable object data on the one hand and improves the attention module to enhance the richness of feature extraction on the other hand, thus enabling the method to better identify the objects. Figure 12(c) also shows some improvement in the case of overlapping multiple objects reducing missed and false detections. The object context is enriched by data enhancement, which improves the detection ability of the model for overlapping objects. From the results, it can be seen that the improved method has a better solution for difficult objects such as small underwater samples and fuzzy objects, and the detection performance is better than other methods. In summary, the method presented does have better capability for underwater object detection.

4.6. More Experiments

In order to verify other model performances , we conduct experiments on different brightnesses, different Gaussian noise and different angle interferences.

4.6.1. Robustness Experiments

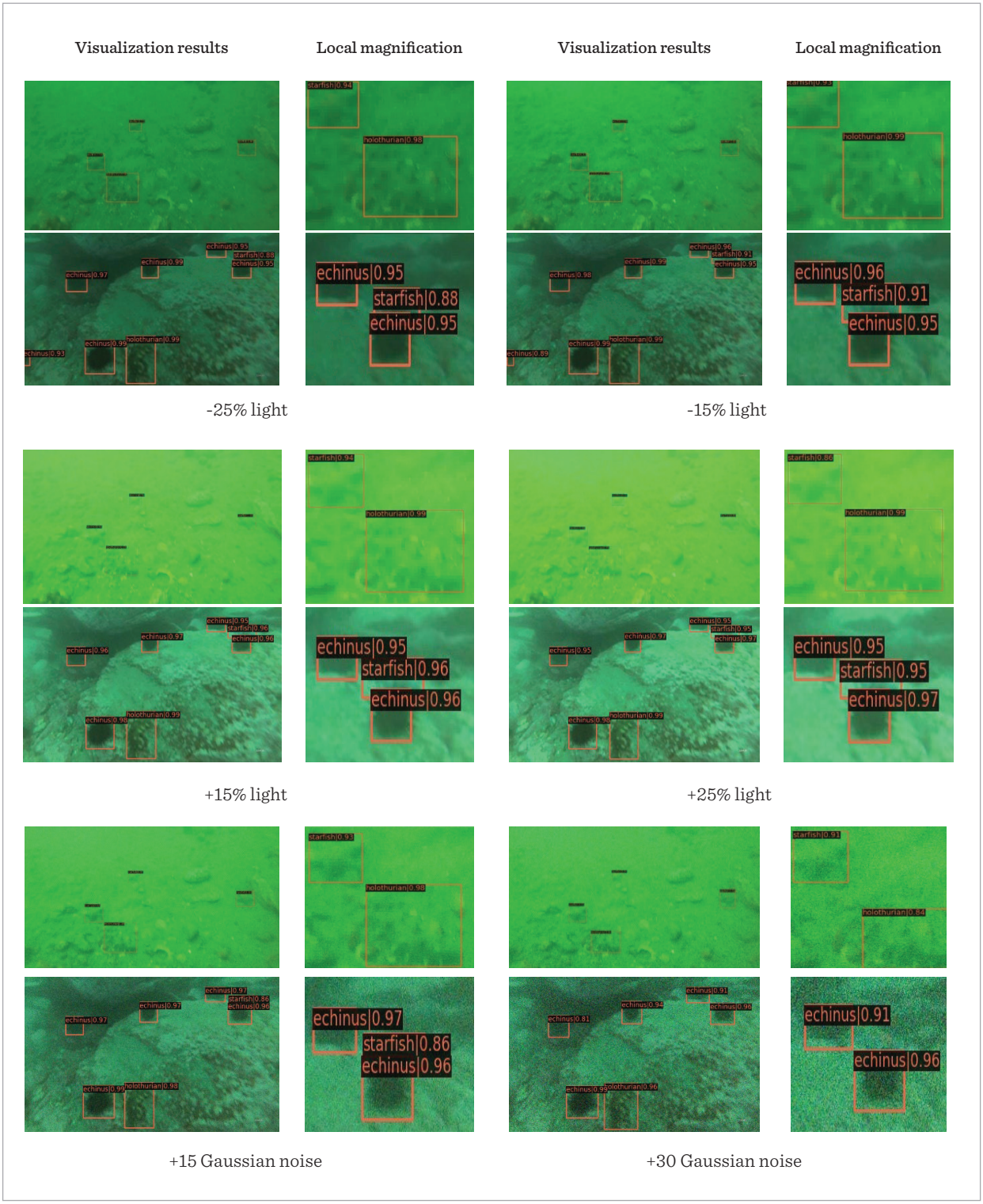
Experiments were performed on the test dataset with different illuminations (±15%, ±25%) and Gaussian noise with variances of 10, 20 and 30. We compare the robustness of the model with YOLOv7, and the average accuracy is listed in Table 8.

As shown in Table 8, although the proposed CDWTN method degrades with the increasing influence of light and Gaussian noise, our method achieves better performance for different noise levels on the URPC dataset. The experimental results of direct testing show that light within 15% has little effect on the image. Gaussian noise with a variance of 10 has less effect on accuracy. These results prove that the robustness of the model is good, which learns various information well in the data. The visualization results with local magnification are shown in Figure 13.

Table 8
Experiments on the robustness of the model

Changes	Ours		YOLOv7	
	AP@0.5	AP@0.50:0.95	AP@0.5	AP@0.50:0.95
None	88.3	56.8	86.9	51.7
-25% light	86.9	53.3	82.3	49.3
-15% light	88.0	54.8	86.0	51.1
+15% light	88.1	55.4	86.3	51.2
+25% light	87.2	53.6	84.2	50.6
+10 Gaussian noise	83.7	50.7	81.8	48.9
+20 Gaussian noise	77.2	43.3	75.6	39.4
+30 Gaussian noise	60.9	34.7	59.1	30.4

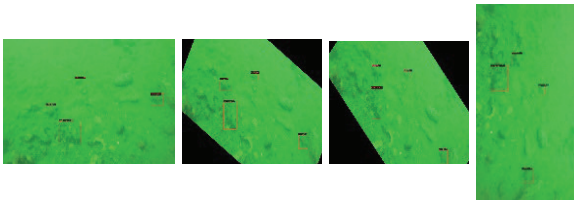
Figure 13
Visualization results of experiments on the robustness



4.6.2. Viewpoint and Size Invariance

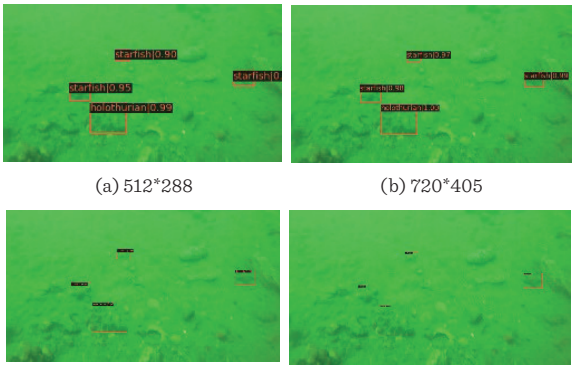
We rotated the image and changed the image size. The detection results can be seen in Figure 14 and Figure 15, which prove the viewpoint and size invariance of the model when the features of the object are not destroyed. For these images of different rotation angles and sizes, our method has the same efficient detection results. This is because multi-scale information has been learned in the CDWTN, and is better integrated under the attentional supervision.

Figure 14
Results of different rotation angles



(a) original (b) 30° (c) 45° (d) 90°

Figure 15
Results of different image sizes



(a) 512*288 (b) 720*405

(c) 1920*1080 (d) 3840*2160

5. Conclusion

In this paper, a new underwater object detection method, CDWTN, is proposed by combining a Transformer with CNN techniques. Specifically, blur prior fusion is proposed to adapt to the blurred scene. A CDW Transformer block is proposed for the precise positioning of underwater objects and effective reduction of the intermediate computation. To address the problem of small object detection, weighted supervision is adopted to integrate multi-scale information in the novel ASFPN. In addition, other adjustments are also made for accuracy. The improved CDWTN method focuses more on the spatial location information of the object while retaining more global and local information. Finally, the experimental results show that the proposed method receives the state-of-the-art results.

In recent years, the rapid development and popularization of artificial intelligence (AI) technology have further enhanced the capabilities of the Internet of Things (IoT), and artificial intelligence technologies such as object detection can be used to post-process the information collected by sensors. In marine-related fields, the IoT is gradually being applied to many marine fields such as ocean observation, island ecological monitoring and intelligent ships. This paper mainly studies the object detection method based on underwater optical images, which can be applied to specific tasks such as fry detection in marine ranching or ship reef detection. Higher precision detection networks can increase the success rate of underwater autonomous operations. In future works, we will further study the proposed model in other image tasks.

Acknowledgement

This work was supported by grants from National Natural Science Foundation of China (NO.62176150).

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S. End-to-end Object Detection with Transformers. European Conference on Computer Vision, 2020, 213-229. https://doi.org/10.1007/978-3-030-58452-8_13
2. Chen, L., Zhou, F., Wang, S., Dong, J., Li, N., Ma, H., Zhou, H. SWIPENET: Object Detection in Noisy Underwater Images. arXiv, 2020, arXiv:2010.10006. https://doi.org/10.48550/arXiv. 2010. 10006
3. Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Lin, D. Hybrid Task Cascade for Instance Segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, 4974-4983. https://doi.org/10.1109/CVPR.2019.00511

4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv, 2020, arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929
5. Guan, T., Li, C., Gu, K., Liu, H., Zheng, Y., Wu, X. J. Visibility and Distortion Measurement for No-reference Dehazed Image Quality Assessment via Complex Contourlet Transform. IEEE Transactions on Multimedia, 2022. https://doi.org/10.1109/TMM.2022.3168438
6. Guo, C., Fan, B., Zhang, Q., Xiang, S., Pan, C. Augfpn: Improving Multi-scale Feature Learning for Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 12595-12604. https://doi.org/10.48550/ arXiv.1912.05384
7. Guo J, Han K, Wu H, et al. Positive-Unlabeled Data Purification in the Wild for Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, 2653-2662. https://doi.org/10.1109/CVPR46437.2021.00268
8. Guo, P., Boyer, F., Chang, X., Hayashi, T., Higuchi, Y., Inaguma, H., Zhang, Y. Recent Developments on Espnet Toolkit Boosted by Conformer. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, 5874-5878. https://doi.org/10.1109/ICASSP39728.2021.9414858
9. Huang, P. X., Boom, B. J., Fisher, R. B. Underwater Live Fish Recognition Using a Balance-guaranteed Optimized Tree. Asian Conference on Computer Vision, 2012, 422-433. https://doi.org/10.1007/978-3-642-37331-2_32
10. Lei, F., Tang, F., Li, S. Underwater Target Detection Algorithm Based on Improved YOLOv5. Journal of Marine Science and Engineering, 2022, 10(3), 310. https://doi.org/10.3390/jmse10030310
11. Li, J., Eustice, R. M., Johnson-Roberson, M. High-level Visual Features for Underwater Place Recognition. IEEE International Conference on Robotics and Automation (ICRA), 2015, 3652-3659. https://doi.org/10.1109/ICRA.2015.7139706
12. Li, X., Shang, M., Qin, H., Chen, L. Fast Accurate Fish Detection and Recognition of Underwater Images with Fast R-cnn. OCEANS, 2015, 1-5. https://doi.org/10.23919/OCEANS.2015. 7404464
13. Li, C., Guan, T., Zheng, Y., Zhong, X., Wu, X., Bovik, A. Blind Image Quality Assessment in the Contourlet Domain. Signal Processing: Image Communication, 2021, 91, 1-11. https://doi.org/10.1016/j.image.2020.116064
14. Li, K., Wang, Y., Zhang, J., Gao, P., Song, G., Liu, Y., Qiao, Y. Uniformer: Unifying Convolution and Self-attention for Visual Recognition. arXiv, 2022, arXiv:2201.09450. https://doi.org/10.48550/arXiv.2201.09450
15. Li, C., Guan, T., Zheng, Y., Jin, B., Wu, X., Bovik, A. Completely Blind Image Quality Assessment via Contourlet Energy Statistics. IET Image Processing, 2021, 15(2), 443-453. https://doi.org/10.1049/ipr2.12034
16. Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S. Feature Pyramid Networks for Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 2117-2125. https://doi.org/10.1109/CVPR.2017.106
17. Lin, W. H., Zhong, J. X., Liu, S., Li, T., Li, G. RoIMix: Proposal-fusion among Multiple Images for Underwater Object Detection. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, 2588-2592. https://doi.org/10.1109/ICASSP40776.2020
18. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J. Path Aggregation Network for Instance Segmentation. Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition, 2018, 8759-8768. https://doi.org/10.1109/CVPR.2018.00913
19. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhang, L. DAB-DETR: Dynamic Anchor Boxes Are Better Queries for DETR. arXiv, 2022. arXiv:2201.12329. https://doi.org/10.48550/arXiv.2201.12329
20. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, 10012-10022. https://doi.org/10.1109/ICCV48922.2021.00986
21. Liu, H., Song, P., Ding, R. Towards Domain Generalization in Underwater Object Detection. IEEE International Conference on Image Processing (ICIP), 2020, 1971-1975. https://doi.org/10.1109/ICIP40778.2020.9191364
22. Liu, C., Li, H., Wang, S., Zhu, M., Wang, D., Fan, X., Wang, Z. A Dataset and Benchmark of Underwater Object Detection for Robot Picking. IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 2021, 1-6. https://doi.org/10.1109/ICMEW53276.2021.9455997
23. Mandal, R., Connolly, R. M., Schlacher, T. A., Stantic, B. Assessing Fish Abundance from Underwater Video Using Deep Neural Networks. International Joint Conference on Neural Networks (IJCNN), 2018, 1-6. https://doi.org/10.1109/IJCNN.2018.8489482

24. Pan, X., Ge, C., Lu, R., Song, S., Chen, G., Huang, Z., Huang, G. On the Integration of Self-attention and Convolution. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 815-825. https://doi.org/10.1109/CVPR52688.2022.00089

25. Pinheiro, P. O., Collobert, R. From Image-level to Pixel-Level Labelling with Convolutional Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, 1713-1721. https://doi.org/10.1109/CVPR.2015.7298780

26. Połap, D., Wawrzyniak, N., Włodarczyk-Sielicka, M.. Side-scan Sonar Analysis Using Roi Analysis and Deep Neural Networks. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60, 1-8. https://doi.org/10.1109/TGRS.2022.3147367

27. Quang, T. N., Lee, S., Song, B. C. Object Detection Using Improved Bi-directional Feature Pyramid Network. Electronics, 2021, 10(6), 746. https://doi.org/10.3390/electronics10060746

28. Santurkar, S., Tsipras, D., Ilyas, A., Madry, A. How does Batch Normalization Help Optimization. Advances in Neural Information Processing Systems, 2018, 31. https://hdl.handle.net/1721.1/137779

29. Tharwat, A., Hemedan, A. A., Hassanien, A. E., Gabel, T. A Biometric-based Model for Fish Species Classification. Fisheries Research, 2018, 204, 324-336. https://doi.org/10.1016/j.fishres.2018.03.008

30. Wang, W., Xie, E., Li, X., Fan, D. P., Song, K., Liang, D., Shao, L. Pyramid Vision Transformer: A versatile Backbone for Dense Prediction without Convolutions. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, 568-578. https://doi.org/10.1109/ICCV48922.2021.00061

31. Wen, F., Ying, R., Liu, Y., Liu, P., Truong, T. K. A Simple Local Minimal Intensity Prior and An Improved Algorithm for Blind Image Deblurring. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(8), 2923-2937. https://doi.org/10.1109/TCSVT.2020.3034137

32. Woo, S., Park, J., Lee, J. Y., Kweon, I. S. Cbam: Convolutional Block Attention Module. Proceedings of the European Conference on Computer Vision (ECCV), 2018, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

33. Yan, J., Zhou, Z., Su, B., Xuanyuan, Z. Underwater Object Detection Algorithm Based On Attention Mechanism and Cross-Stage Partial Fast Spatial Pyramidal Pooling. Frontiers in Marine Science, 2022, 2299. https://doi.org/10.3389/fmars.2022.1056300 https://doi.org/10.3389/fmars.2022.1056300

34. Zhang, M., Xu, S., Song, W., He, Q., Wei, Q. Lightweight Underwater Object Detection Based on YOLOv4 and Multi-scale Attentional Feature Fusion. Remote Sensing, 2021, 13(22), 4706. https://doi.org/10.3390/rs13224706

35. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J. Deformable detr: Deformable Transformers for End-to-end Object Detection. arXiv, 2020, arXiv:2010.04159. https://doi.org/10.48550/arXiv.2010.04159

This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 (CC BY 4.0) License (http://creativecommons.org/licenses/by/4.0/).