Hindawi
Journal of Sensors
Volume 2024, Article ID 9796471, 1 page
https://doi.org/10.1155/2024/9796471

Retraction

Retracted: A Novel Attention-Based Lightweight Network for Multiscale Object Detection in Underwater Images

Journal of Sensors

Received 23 January 2024; Accepted 23 January 2024; Published 24 January 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

In addition, our investigation has also shown that one or more of the following human-subject reporting requirements has not been met in this article: ethical approval by an Institutional Review Board (IRB) committee or equivalent, patient/participant consent to participate, and/or agreement to publish patient/participant details (where relevant).

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] J. Wang, X. He, F. Shao et al., "A Novel Attention-Based Lightweight Network for Multiscale Object Detection in Underwater Images," *Journal of Sensors*, vol. 2022, Article ID 2582687, 14 pages, 2022.

Hindawi
Journal of Sensors
Volume 2022, Article ID 2582687, 14 pages
https://doi.org/10.1155/2022/2582687

Research Article

A Novel Attention-Based Lightweight Network for Multiscale Object Detection in Underwater Images

Jinkang Wang, Xiaohui He, Faming Shao, Guanlin Lu, Qunyan Jiang, Ruizhe Hu, and Jinxin Li

Department of Mechanical Engineering, College of Field Engineering, PLA Army Engineering University, Nanjing 210007, China

Correspondence should be addressed to Xiaohui He; gcbhxh@aeu.edu.cn

Received 11 July 2022; Accepted 14 August 2022; Published 7 September 2022

Academic Editor: Yuan Li

Underwater images have low quality, and underwater targets have different sizes. The mainstream target detection networks cannot achieve good results in detecting objects from underwater images. In this study, a lightweight underwater multiscale target detection model with an attention mechanism is designed to solve the above problems. In this model, MobileNetv3 is used as the backbone network for preliminary feature extraction. The lightweight feature extraction module (LFEM) pays attention to the feature map at the channel and space levels. The features with large weights are promoted, while the features with small weights are suppressed. Meanwhile, cross-group information exchange enriches the semantic information and location information of the objects. The context aggregation module (CIAM) pools the extracted feature maps to obtain feature pyramids, and it uses the upsampling-feature refinement-cascade addition (URC) method to effectively fuse global context information and enhance the feature representation. The scale normalization for feature pyramids (SNFP) performs adaptive multiscale perception and multianchor detection on feature maps to cover objects of different sizes and realize multiscale object detection in underwater images. The proposed network can realize lightweight feature extraction, effectively handle the global relationship between the underwater scene and the object while expanding the receptive field, traverse the objects of different scales, and achieve adaptive multianchor detection of multiscale objects in underwater images. The experimental results indicate that our method achieves an average accuracy of 81.94% and a detection speed of 44.3 FPS on a composite dataset. Also, our method is better than the mainstream object detection networks in terms of detection accuracy, lightweight design, and real-time performance.

1. Introduction

With the rapid growth of the world population and the increasing shortage of available inland resources, the rich biological and mineral resources in the ocean become important for human survival in the future. In the process of ocean exploration and research, underwater object detection from underwater images plays an important role in underwater applications such as military operations, resource exploration, environmental protection, and biological research.

Underwater object detection can be combined with an underwater robot to monitor and search the interested targets with the assistance of the underwater camera, which has important research value and application prospects. As a branch of computer vision, underwater object detection based on optical images has become a new research field in ocean exploration.

In the complex imaging environment, the quality of underwater images taken by underwater cameras deteriorates due to factors such as illumination, medium, wavelength, and vibration [1]. This has a great influence on the accuracy of target detection. Underwater objects have various scales, and the semantic information of large-scale objects is in deep feature maps. However, the detailed information of small-scale objects will gradually decrease or even be lost during the downsampling process. This makes the task of underwater image object detection more difficult. The existing methods improve the detection effect of multiscale objects by fusing features and constructing complex networks, which improves the detection accuracy at the cost of detection efficiency. Meanwhile, the real-time performance of underwater object detection is greatly reduced. Improving the detection efficiency while improving the detection accuracy is an urgent problem to be solved in underwater object detection.

Aiming at the above problems, this paper proposes an attention-based lightweight model for multiscale object detection. The lightweight feature extraction module (LFEM) adopts dual attention to pay attention to the feature map at the channel level [2] and spatial level [3], and it uses "channel shuffle" [4] to exchange information across groups to enrich semantic information of multiscale objects. The context aggregation module (CIAM) uses different scales of pooling to obtain feature pyramids, and it adopts the original upsampling-feature refinement-cascade addition module (URC) to obtain both global semantic information and local detail information. The scale normalization for the feature pyramid (SNFP) module performs adaptive multiscale perception and multianchor detection on feature maps of different sizes to realize multiscale object detection in underwater images. Experimental results show that our proposed method outperforms current mainstream methods in terms of average accuracy, speed, and resource consumption.

The contributions of this paper are summarized as follows:

(1) Aiming at the problems faced by underwater image object detection, a lightweight feature extraction module is proposed, which can effectively extract feature-layer information while reducing model parameters and improving detection efficiency

(2) In the CIAM module, the strategy of "upsampling-feature refinement-cascade addition" is proposed to increase the receptive field and improve the network's ability to obtain global context information

(3) To obtain a better detection effect, SNFP is proposed to perform adaptive multiscale perception and multianchor detection of different scales

(4) The experimental results show that our proposed network on the datasets RUIE, HabCam UID, and SQUID achieves better performance than the current mainstream methods

2. Related Works

The current object detection technology is very mature, and underwater image target detection is developing rapidly as a new branch of object detection. Balancing detection accuracy and speed is a research hotspot of underwater object detection [5]. The quality of underwater images is seriously degraded, and the size of underwater targets varies greatly. In addition, underwater object detection has relatively high real-time requirements. How to perform accurate, fast, and stable detection of multiscale targets in complex underwater scenes is worth studying.

2.1. Object Detection. According to the presence or absence of candidate frame generation stages, object detection methods based on deep learning can be divided into two-stage object detection methods and single-stage object detection methods. The two-stage object detection methods, such as R-CNN [6], Fast R-CNN [7], and Faster R-CNN [8], first extract candidate regions and then perform secondary correction based on the candidate regions to obtain the detection results. The detection accuracy is high, but the detection speed is slow due to a large number of convolution operations. The single-stage object detection methods, such as SSD [9] and YOLO series [10–13], do not need to extract candidate frames, which directly calculate the images to generate the detection results. The detection speed is fast, but the detection accuracy is low. Some researchers combined the two types of methods to balance detection accuracy and speed. RON [14] is an efficient and general object detection model proposed based on SSD and Faster R-CNN. The experimental results indicate that RON achieves much higher detection accuracy than SSD under the same condition, and the detection speed is three times faster than that of faster R-CNN. RefineDet [15] integrates RPN, FPN [16], and SSD algorithms, which improves the detection accuracy on the PASCAL VOC 2007 dataset [17] to 80.0% while maintaining the efficiency of SSD. RetinaNet [18] combines FPN and FCN networks and adopts an improved cross-entropy focal loss to effectively eliminate the problem of class imbalance. STDN [19] proposes a scale-transfer layer to generate large-scale feature maps without increasing the number of parameters and computation amount, which improves the detection efficiency.

In recent years, the field of underwater image object detection based on deep learning has also developed rapidly. Chen et al. [20] designed SWIPENet to detect underwater small-sample objects. SWIPENet uses a sample reweighting algorithm IMA and introduces a dilated convolutional layer to obtain a large receptive area without sacrificing the resolution of the feature map. Lin et al. [21] proposed an image enhancement method based on candidate box fusion to generate training samples that simulate overlapping, occlusion, and blurring, which improves the mean average precision (mAP) and robustness of the model. Zheng et al. [22] first enhanced the image for better contrast and then separated objects and backgrounds to improve object detection performance. Zeng et al. [23] proposed Faster R-CNN-AON, in which the Faster R-CNN network and the AON [24] network compete and learn together so that the detection network can obtain better robustness, which effectively prevents the detection network from overfitting and greatly improves the detection accuracy.

2.2. Lightweight Module. The deep object detection network usually contains a large number of parameters, which requires huge storage space and running space to complete the detection task. To migrate the underwater image object detection algorithm from the server to the mobile terminal, it is urgent to lightweight the object detection model.

MobileNetv1 [25] divides the convolution of the standard object detection network structure into a depth-wise separable convolution and a point-wise convolution, which reduces the network weight parameters and the model calculation amount and improves the calculation speed. MobileNetv2 [26] uses linear bottlenecks to remove the nonlinear activation layer behind the small-dimensional output layer and adopts the inverted residual strategy, which greatly improves the model effect. Based on the combination of the depth-wise separable convolution of MobileNetv1 and the linear bottleneck and inverse residual structure of MobileNetv2, MobileNetv3 [27] introduces the SE attention module and updates the activation function to make the convolutional neural network more lightweight. ShuffleNet v2 [28] uses the channel shuffle method to shuffle the order of each feature map to form a new feature map to achieve cross-group information exchange. Ghostnet [29] uses simple linear operations to obtain redundant feature maps to enhance features and increase channels, which greatly reduces the computation amount and improves computational efficiency.

Lightweight models are common in conventional object detection, but there are few studies on underwater image object detection. This study combines the characteristics of different lightweight models and transforms them. Meanwhile, a lightweight feature extraction module is proposed to improve the real-time performance of underwater image object detection.

2.3. Multiscale Fusion. The scale problem of object detection always affects the detection effect, and the accuracy of detecting extremely large or small objects will be significantly reduced. Many effective network frameworks have been designed for multiscale detection.

The image pyramid scales images at different scales, randomly trains images of different scales, and forces the neural network to adapt to objects of different scales, which preliminary improves detection results. SNIP [30] achieves selective training by selectively returning gradients, reducing the impact of domain shift, and achieving better detection results for objects of extreme sizes. Based on SNIP, SNIPER [31] only processes context regions around ground-truth instances on the image pyramid, and the training speed is increased by three times. FPN [16] upsamples each layer from top to bottom, and it combines high-level features of deep convolutional layers with low-level features of shallow convolutional layers to obtain more accurate pixel position information; PANet [32] creates a bottom-up feature refusion side path based on FPN and reconstructs a pyramid that strengthens spatial information, which makes full use of the information of each feature layer; the SPP [33] module adopts the multiscale block method of SPM [34] and performs pooling operations on each block to convert the feature maps of any size into a fixed-length feature vector. ASPP [35] uses atrous convolution to build convolution kernels with different receptive fields to obtain rich multiscale object information. To simulate the receptive field structure of the human visual system as much as possible, RFBNet [36] integrates the characteristics of the Inception module [37] and the ASPP module. This greatly improves the accuracy while ensuring the detection speed.

Underwater images not only have large differences in object size but also have a large number of small objects. Comprehensively considering detection speed and accuracy, this paper proposes SNFP, which combines the advantages of SNIP and FPN and performs adaptive multiscale perception and multianchor detection of different scales.

3. Overview of Recommended Methods

To solve the difficulties encountered in the process of underwater image multiscale object detection, this paper proposes a new lightweight object detection network, and the algorithm flow is shown in Figure 1. First, the original underwater image is preliminarily extracted by MobileNetv3. Then, LFEM pays attention to the feature map at the channel and spatial levels, respectively, and it realizes cross-group communication of the feature information through channel shuffle. Next, CIAM pools the extracted feature maps to obtain feature pyramids, and it fuses feature maps of different scales using the original URC method to effectively fuse global context information and enhance feature representation ability. Finally, the SNFP performs adaptive multiscale perception and multianchor detection on feature maps of different sizes to cover objects of different sizes and realize multiscale object detection in underwater images. According to the characteristics of underwater images, our proposed network achieves lightweight feature extraction, effectively handles the global relationship between the scene and the objects while expanding the receptive field, and performs adaptive multianchor box detection for objects with large-scale differences. Based on this, the proposed method can effectively detect multiscale objects in different water scenes.

3.1. Lightweight Feature Extraction Module. The traditional feature extraction network usually consists of a large number of convolutions, which consumes huge computing resources and has poor real-time detection performance. To avoid this problem, this paper designs a lightweight feature extraction module for underwater images, and its structure is shown in Figure 2.

3.1.1. Depth-Wise Separable Convolution and Point-Wise Convolution. Depth-wise separable convolution splits convolution kernel into single channel form and convolutes each channel without changing the depth of the feature map. Point-wise convolution uses a 1 × 1 convolution kernel to fuse the feature maps obtained in depth-wise separable convolution to solve the problem of unsmooth information exchange between feature maps. In depth-wise separable convolution, one convolution kernel is only responsible for one channel. Assume that there are M input features, N output features, the input feature size is D_F, and M convolution kernels of $D_k \times D_k$ are required. To output N characteristic maps, point-wise convolution uses $N 1 \times 1$ convolution kernels for convolution. The ratio of the amount of calculation with the standard convolution is

$$\frac{D_k \cdot D_k \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F}{D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F} = \frac{1}{N} + \frac{1}{{D_K}^2}. \quad (1)$$

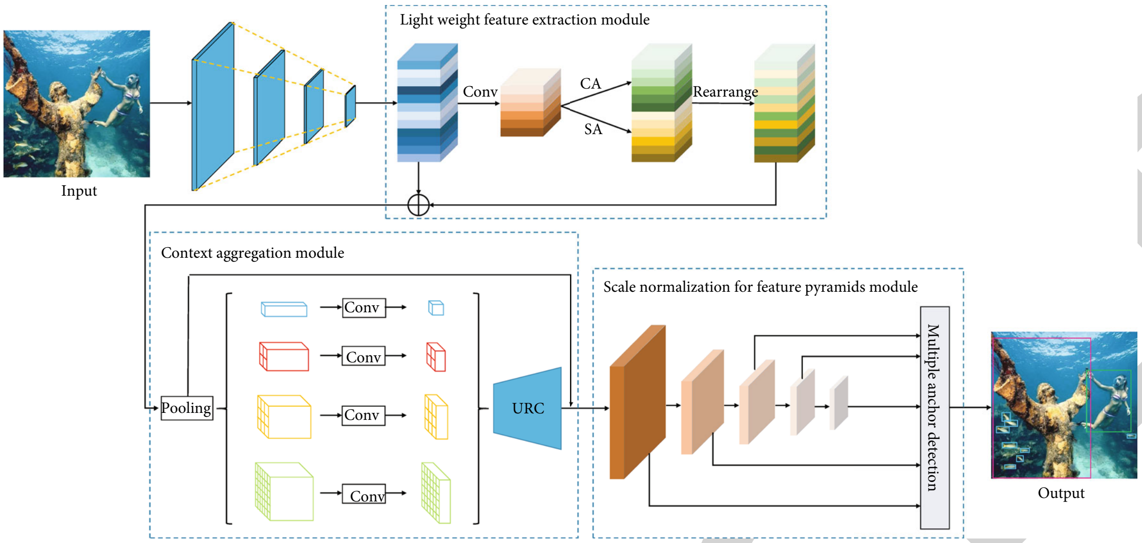

Input
Light weight feature extraction module
Conv
CA
SA
Rearrange
Context aggregation module
Pooling
Conv
Conv
Conv
Conv
URC
Scale normalization for feature pyramids module
Multiple anchor detection
Output

Figure 1: The pipelines of our method.

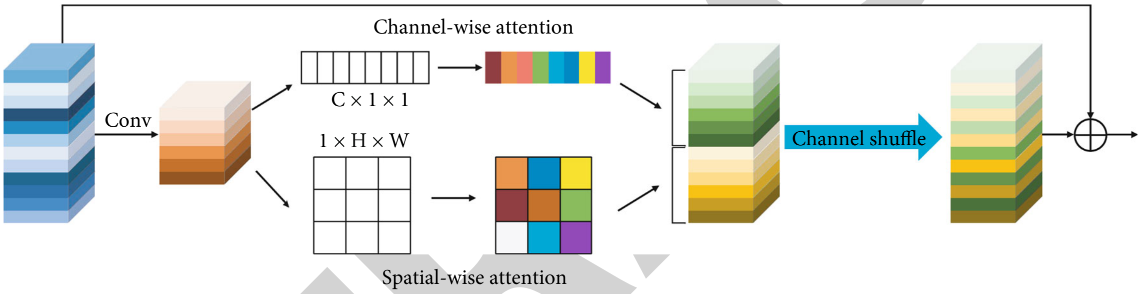


Figure 2: The structure of the lightweight feature extraction module.

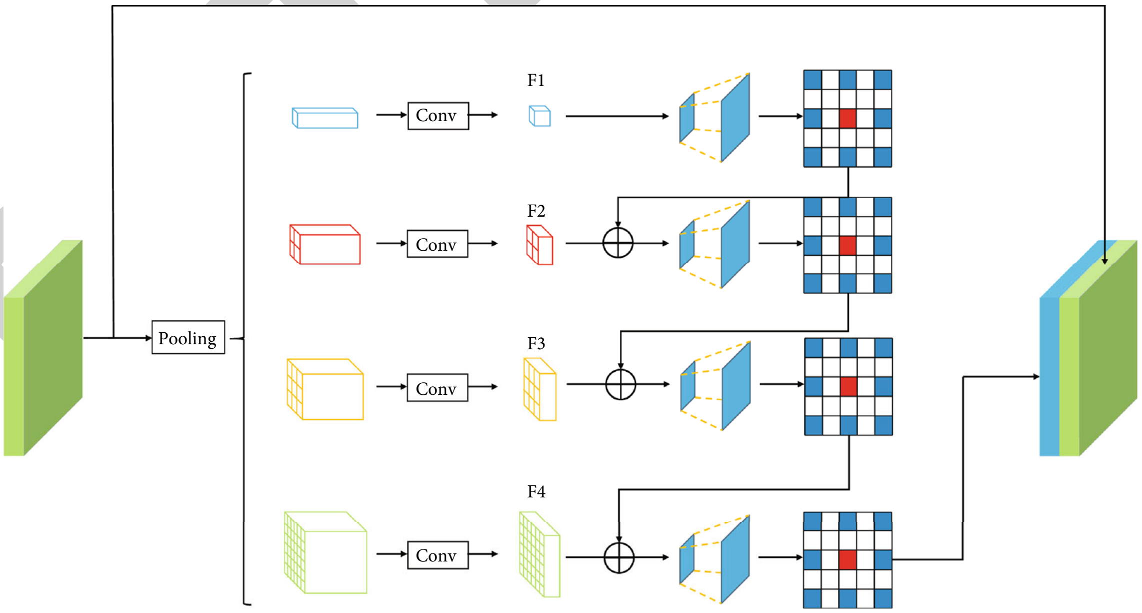


Figure 3: The structure of CIAM.

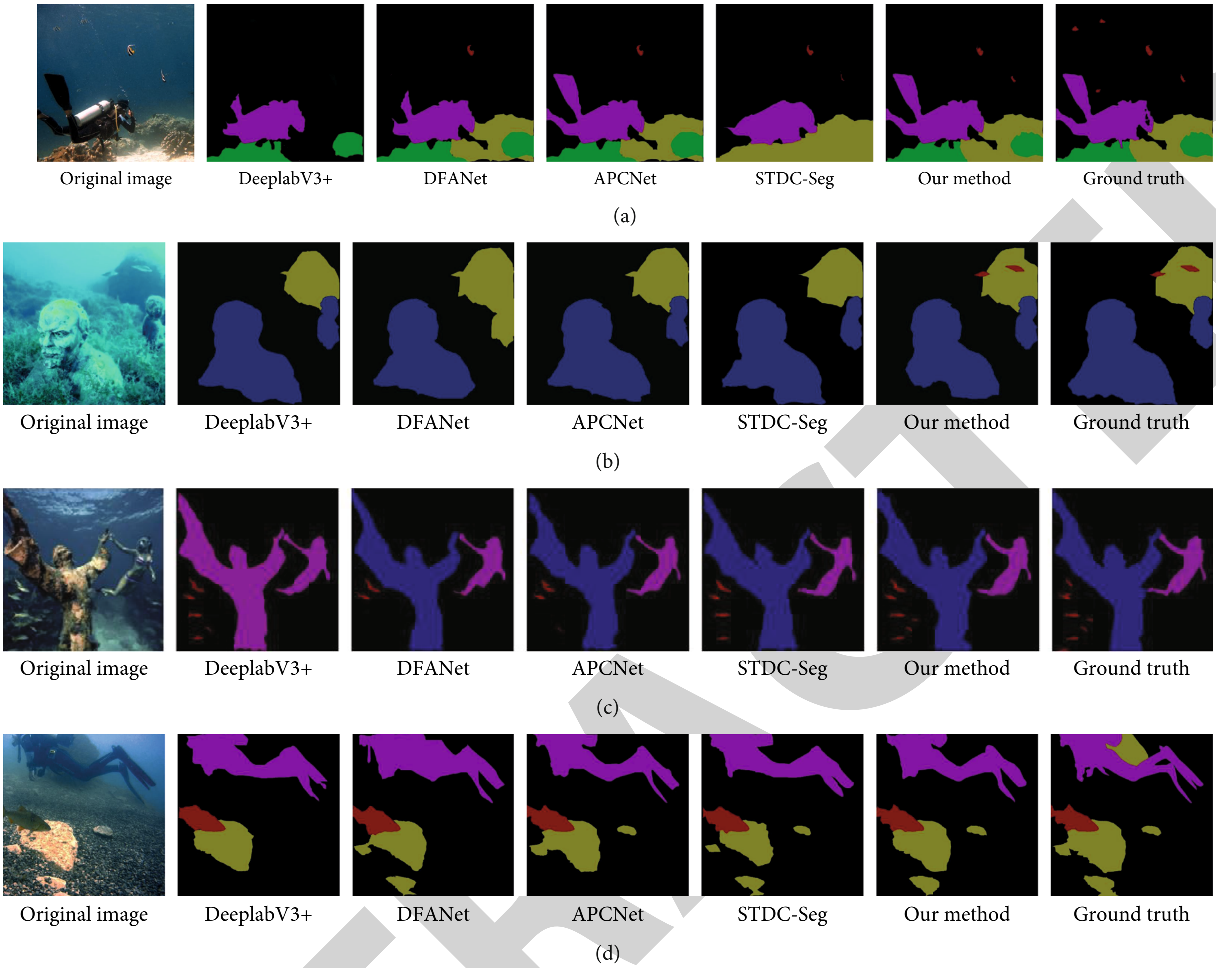


Figure 4: Qualitative comparisons with the four most advanced classical segmentation methods. From the left to right are the original images, the enhanced images, and the results generated by Deeplab V3+, DFANet, APCNet, STDC-Seg, CIAM, and the ground truth.

The ratio of the number of parameters is

$$\frac{D_k \cdot D_k \cdot M + M \cdot N}{D_k \cdot D_k \cdot M \cdot N} = \frac{1}{N} + \frac{1}{{D_K}^2}. \tag{2}$$

Compared with general convolution, when D_k and N are large, depth-wise separable convolution and point-wise convolution have great advantages in terms of parameter size and calculation speed.

3.1.2. Double Attention Mechanism. The parallel dual attention mechanism extracts and retains key information. The channel attention network captures channels containing important object feature information and assigns large weight values to these channels. The feature map is compressed by global pooling to generate a C-dimensional feature vector, which is then processed by the full connection layer f_{CA}. The feature vector is mapped to the range of [0, 1] by a sigmoid gate function, and weighting operations are performed finally. The calculation process is shown in

$$\mathrm{CA} = F(v, W) = \sigma_1(fc_2(\delta(fc_1(v, W_1)), W_2)), \tag{3}$$

$$\tilde{f}_{\mathrm{CA}} = \mathrm{CA} \times f_{\mathrm{CA}}, \tag{4}$$

where W represents the weight parameter that needs to be updated, v represents the C-dimension feature vector, σ_1 represents the sigmoid activation operation, fc represents the fully connected layer, δ represents the Relu activation function, and $\tilde{f}_{\mathrm{CA}}$ represents the weighted feature map.

The function of spatial attention is to capture local regions in feature maps that contain important detail information. The feature map is passed through two parallel asymmetric convolutional layers, and the output is added along the channel direction. Finally, the feature values are mapped to the range of $[0, 1]$ by the sigmoid gate function, and then, weighting operations are performed. The calculation process is shown in

$$C_1 = \mathrm{conv}_2\left(\mathrm{conv}_1\left(Y, U_1^1\right), U_1^2\right), \tag{5}$$

$$C_2 = \mathrm{conv}_1\left(\mathrm{conv}_2\left(Y, U_2^1\right), U_2^2\right), \tag{6}$$

$$\mathrm{SA} = F(Y, U) = \sigma_2(C_1 + C_2), \tag{7}$$

$$\tilde{f}_{\mathrm{SA}} = \mathrm{SA} \times f_{\mathrm{SA}}, \tag{8}$$

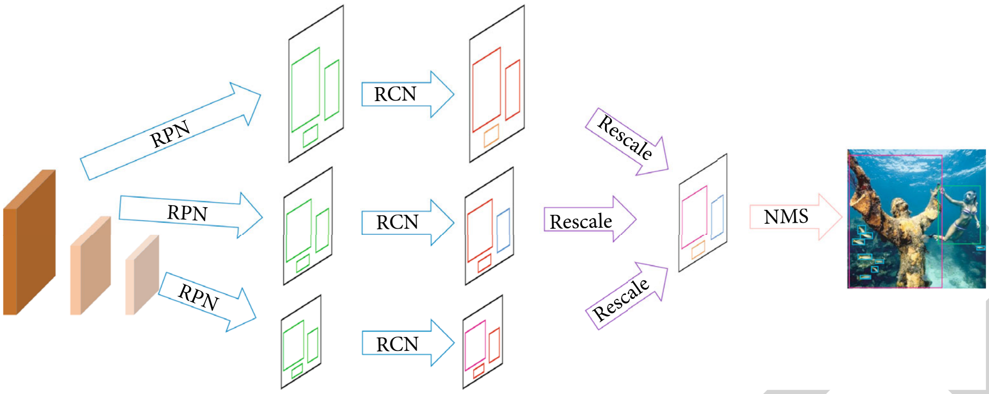


Figure 5: The structure of SNFP.

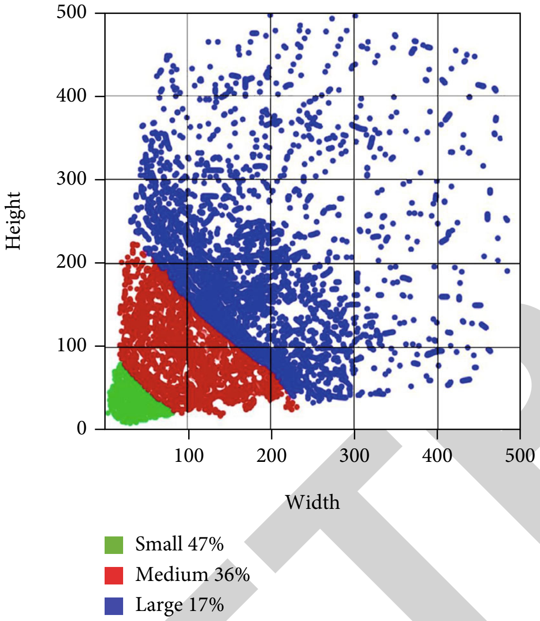


Figure 6: Instance size distribution of the CUID dataset.

where U represents the convolution kernel parameter, conv$_1$ and conv$_2$ represent the asymmetric convolution layer, respectively, Y represents the input feature map, σ_2 represents the sigmoid activation operation, and $\tilde{f}_{SA}$ represents the weighted feature map.

In general, channel attention focuses on "what" is an effective feature that requires specific attention, and spatial attention focuses on "which" is the most informative feature. The dual attention mechanism can purify the features adaptively while extracting and retaining key features.

3.1.3. Channel Shuffle. As shown in Figure 2, channel shuffle is used to rearrange the feature maps generated by the two attention networks to realize cross-group information exchange and form a complete feature map of the same size as the original feature map. Cross-group information exchange makes feature extraction more sufficient and greatly improves the feature utilization efficiency of small-scale objects.

3.2. Context Information Aggregation Module. For underwater images, low resolution leads to unclear feature expression. Under the layer-by-layer convolution, the details of the feature map are missing, and the correlation between pixels is gradually weakened, which makes it difficult to obtain scene context information. To aggregate the context information of different areas and improve the ability of the network to obtain global information, this paper designs the context information aggregation module, as shown in Figure 3. The original feature map is pooled with different scales to obtain the feature pyramid. Then, the feature maps with different scales are fused by the URC module to consider global semantic information and local detail information and enhance feature representation ability.

The context information aggregation module uses a PPM-like method to obtain feature maps of different sizes. The input feature map size is 6 × 6, and it is pooled by 6 × 6, 3 × 3, 2 × 2, and 1 × 1 to obtain feature maps with the output sizes of 1 × 1, 2 × 2, 3 × 3, and 6 × 6, respectively. These feature maps of different sizes contain context information of different areas. As shown in Figure 3, feature map F1 is upsampled by bilinear interpolation to increase the resolution. Then, the feature map with increased resolution is refined by atrous convolution with a rate of 2 and added with the feature map F2 pixel-by-pixel to complete the first information fusion between feature maps. The above operation is repeated until the feature map F4 is upsampled to the original feature map size. Subsequently, the output feature map and the original feature map are spliced in the channel dimension, which not only increases the receptive field but also greatly improves the ability of the network to obtain global context information. Finally, the context information aggregation module merges the deep semantic information with the shallow edge line, shape position, and other detailed information, which helps to capture clear object boundary information, refine segmentation results and effectively improve object segmentation accuracy.

To intuitively show the effectiveness of CIAM, the comparison results with the four most advanced classical segmentation methods are presented in Figure 4. From the left to right are the original images, the enhanced images, and the results generated by Deeplab V3+ [38], DFANet [39], APCNet [40], STDC-Seg [41], our method, and the

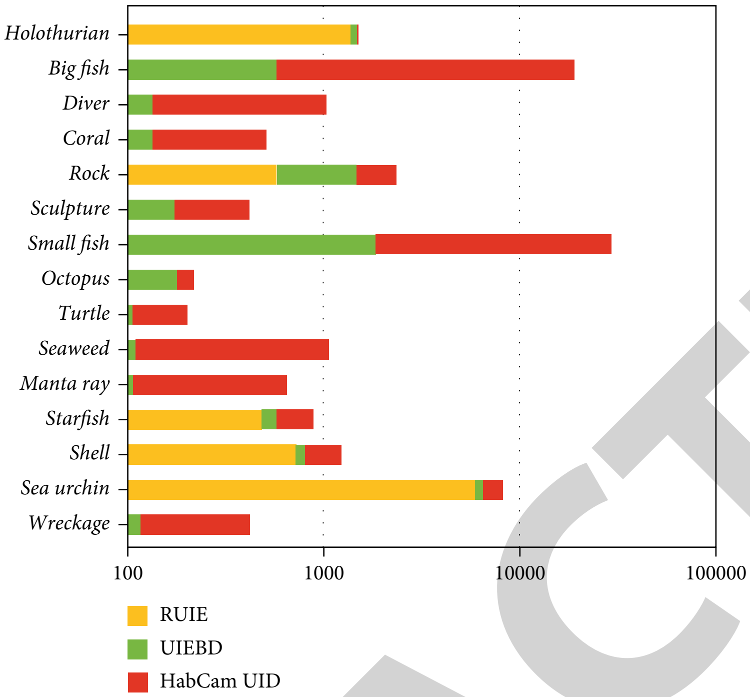


Figure 7: Instance number distribution of the CUID dataset.

Table 1: The experimental environment configuration.

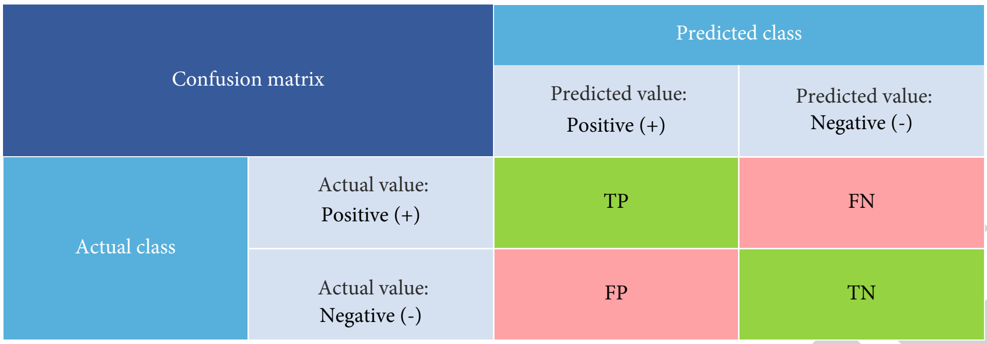
Environment	Version model
Operating system	Windows 10 64-bit
CPU	Intel i7-6700U 4.00 GHz
GPU	NVIDIA RTX 3090 Ti
CUDA	V10.1
PyTorch	V1.5.0
Python	V3.6

ground truth. It can be seen from the experimental results that the proposed context information aggregation module performs the best in terms of segmentation integrity, positioning accuracy, and boundary definition and details, which will contribute to better underwater target detection performance.

3.3. SNFP. Aiming at the difficulty of multiscale object detection in underwater images, this paper designs SNFP for adaptive multiscale prediction and multianchor detection of objects of different scales. Firstly, RPN extracts candidate regions for feature maps of different layers. For a large-scale feature map, the corresponding RPN is only responsible for predicting the magnified small objects, and the original large objects are no longer in the effective range because they are too big. For the small-scale feature map, the corresponding RPN is only responsible for predicting the decreased large objects, and the original small objects are no longer in the effective range because they are too small. RCN extracts anchor frames of different scales on feature layers of different scales, and it displays all the anchor frames on the normalized feature map. Finally, the object detection result is output through nonmaximum suppression, as shown in Figure 5.

4. Experimental Analysis

4.1. Dataset. Experiment is evaluated on three public datasets: RUIE [42], HabCam UID [43], and UIEBD [44]. RUIE is a self-made dataset of the Dalian University of Science and Technology. It consists of 4000 low-resolution underwater images, including underwater targets such as scallop, holothurian, and sea urchin. The HabCam UID dataset is produced by CVPR AAMVEM studio, which consists of 10,465 underwater images with a resolution of 2720×1024 . It contains over 100,000 instances of underwater objects such as fish, scallop, rock, manta ray, and turtle, which is the largest and most diversified underwater image dataset recently released for target detection. The UIEBD dataset contains 950 underwater images of various multiresolution underwater scenes, including diver, sculpture, and other marine life. The three datasets are merged by using the resize operation to cluster the pixels of the large-resolution images and interpolate the pixels of the small-resolution images. To some extent, the image information is extracted, and the pixels are rearranged to the resolution of 512×512. The merged dataset is called CUID (Composite Underwater Image Dataset), and the ratio of the training set to the testing set is 4 : 1. The instance sizes of the CUID dataset are counted, as shown in Figure 6. The small object pixel is within 100×100, the medium object pixel is between 100×100 and 300×300, and the large object pixel is larger than 300×300. The number distribution of each type of object is shown in Figure 7.



Confusion matrix		Predicted class	
		Predicted value: Positive (+)	Predicted value: Negative (-)
Actual class	Actual value: Positive (+)	TP	FN
	Actual value: Negative (-)	FP	TN

Figure 8: The confusion matrix.

Table 2: Comparison of AP values of different methods. The short names are defined as HO—holothurian, BF—big fish, DI—diver, CO—coral, RO—rock, SC—sculpture, SF—small fish, OC—octopus, TU—turtle, SW—seaweed, MR—manta ray, SF—starfish, SH—shell, SU—sea urchin, and WR—wreckage. The best results are marked in bold.

Method	mAP	HO	BF	DI	CO	RO	SC	SF	OC	TU	SW	MR	SF	SH	SU	WR
Yolov5	71.27	72.80	75.54	76.11	63.92	64.54	63.57	69.41	72.61	73.36	64.39	75.49	70.53	69.34	70.02	67.37
RON	74.26	71.93	78.32	77.29	65.47	67.18	68.95	72.87	74.49	75.02	69.71	78.32	73.68	74.33	74.16	70.12
RefineDet	75.06	69.17	77.09	78.31	68.71	67.34	70.13	76.29	76.38	75.85	70.24	78.37	74.92	75.06	75.13	73.21
STDN	76.67	73.28	80.14	77.42	69.18	67.99	69.71	78.51	77.89	76.91	74.06	77.91	75.03	74.91	76.01	74.88
SWIPENet	78.91	70.35	80.59	80.58	70.26	69.43	73.55	**79.48**	78.01	78.25	74.57	79.38	75.86	79.08	75.64	**75.61**
Faster R-CNN-AON	79.68	73.01	**84.47**	**86.03**	71.25	70.05	**73.69**	78.52	79.62	78.84	**75.29**	79.94	78.61	76.31	78.35	75.36
RFBNet	73.99	70.67	83.92	82.63	69.04	67.36	70.19	76.13	74.17	73.90	70.37	74.06	73.31	70.91	72.05	70.91
Ours	**81.94**	**78.33**	83.98	85.14	**73.38**	**72.01**	73.30	78.57	**80.85**	**82.41**	75.17	**81.54**	**81.02**	**79.63**	**80.28**	74.23

Table 3: Comparison of AP values of different methods for objects of different sizes. The best results are marked in bold.

Method	AP_S	AP_M	AP_L	AP_{50}	AP_{75}
Yolov5	40.07	68.31	75.38	62.34	43.64
RON	43.57	72.45	78.03	65.17	45.49
RefineDet	44.21	73.49	79.07	65.42	46.03
STDN	43.61	74.08	80.15	66.39	46.55
SWIPENet	45.82	75.63	81.43	68.21	48.08
Faster R-CNN-AON	46.23	**77.26**	81.59	69.00	49.20
RFBNet	42.48	72.52	77.24	64.38	44.61
Ours	**48.73**	76.90	**83.41**	**69.84**	**49.94**

4.2. Experimental Setting. Our experimental environment is shown in Table 1. The experiment was conducted on a computer equipped with Intel Core i7-6700U @ 4.00 GHz, NVIDIA GeForce RTX 3090 Ti, 8 GB DDR3 memory, and running Windows 10 64-bit operating system. Experiments were implemented on the PyTorch software. The version of CUDA is 10.1, the version of PyTorch is 1.5.0, and the version of Python is v3.6. Our method is accelerated on GPU.

The network uses the SGD [60] optimization strategy with a momentum parameter of 0.95. The learning rate was set to 0.0001, and it then dropped evenly to 0.00001. The batch size was set to 32, the confidence threshold was set to 0.5, and the IOU threshold was set to 0.4. Besides, the dropout was set to 0.5 to prevent overfitting, and the number of training iterations of CUOID was set to 200,000 times.

4.3. Evaluation Indicators. This study adopts AP and mAP as evaluation indicators. The ground truth is obtained through manual annotation. The confusion matrix is shown in Figure 8.

Recall is the ratio of true-positive samples to the sum of true-positive samples and false-negative samples, and its calculation is shown in formula (9). Precision is the ratio of true-positive samples to the sum of true-positive samples and false-positive samples, and its calculation is shown in formula (10).

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{9}$$

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{10}$$

AP represents the model's average detection accuracy for a specific class of objects, and mAP is the average value of AP values under all categories. Their calculations are shown in

$$AP = \int_0^1 P(r)dr, \tag{11}$$

Table 4: Comparison of the model size and real-time performance of different methods.

Method	Backbone	Input size	Parameters (M)	Model size (MB)	FLOPs (G)	FPS
Yolov5	Darknet53	416 × 416	5.4	22.8	40.57	72
RON	VGG16	384 × 384	31.9	128.6	15.47	15
RefineDet	VGG16	320 × 320	50.5	200.2	18.98	40.7
STDN	DenseNet169	513 × 513	29.5	120.1	3.41	28.6
SWIPENet	VGG16	512 × 512	~	~	~	30
Faster R-CNN-AON	VGG16	1000 × 600	84.1	336.3	23.67	24
RFBNet	VGG16	300 × 300	34.5	140.0	45.42	83
Ours	MobileNetv3	512 × 512	7.8	31.2	21.70	44.3

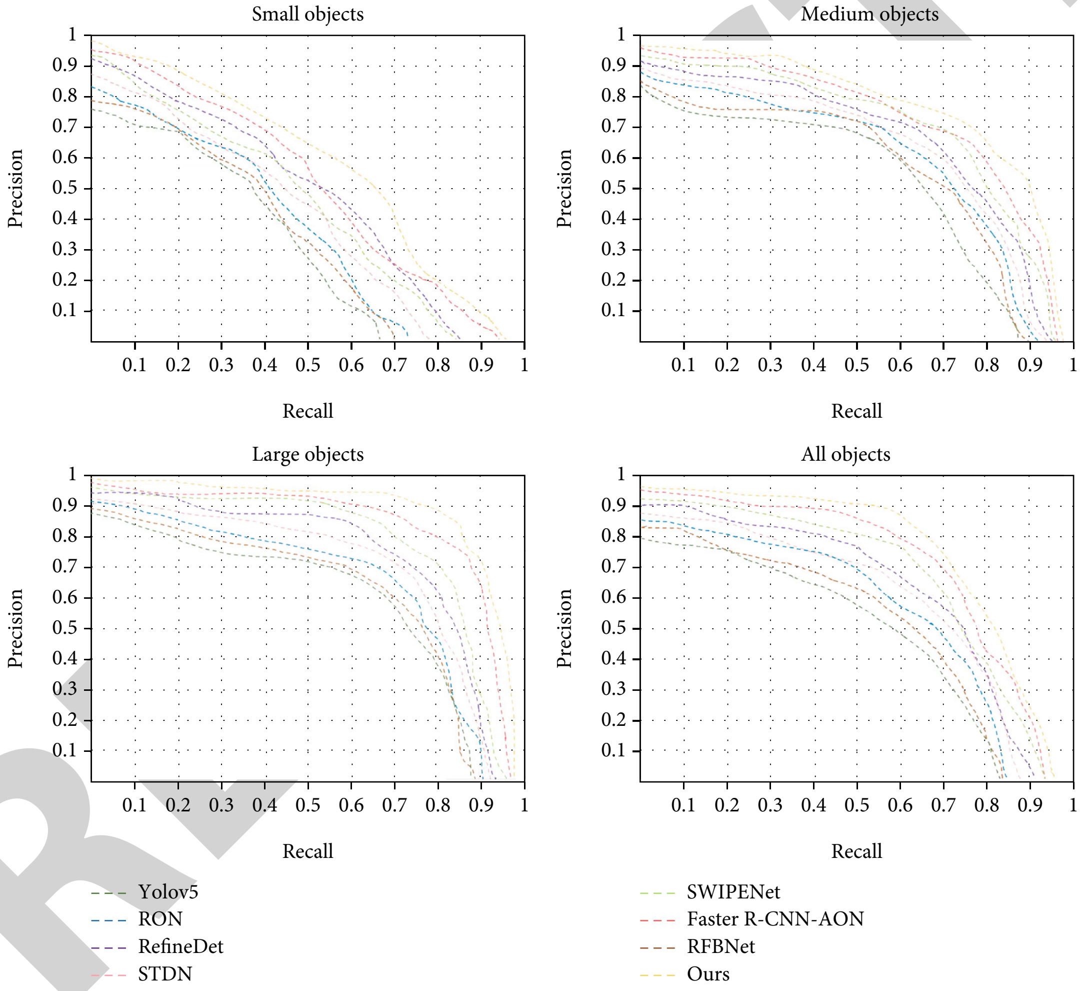


Figure 9: Comparison of the P-R curves of detecting small, medium, large, and all objects on the CUID dataset.

$$\text{mAP} = \frac{1}{Q}\sum_{q=1}^{Q}\text{AP}(q), \tag{12}$$

where $P(r)$ represents the precision value on the P-R curve and AP is the integral calculation of the P-R curve. q represents a specific object class, and Q represents the number of object classes.

4.4. *Experimental*

4.4.1. Objective Evaluation. Table 2 shows the comparison of detection results on Yolov5, RON, RefineDet, STDN, SWIPENet, Faster R-CNN-AON, RFBNet, and our proposed method on the CUID dataset. From Table 2, it can be seen that our method achieves the highest AP value in the detection of objects such as holothurian, coral, rock, and octopus.

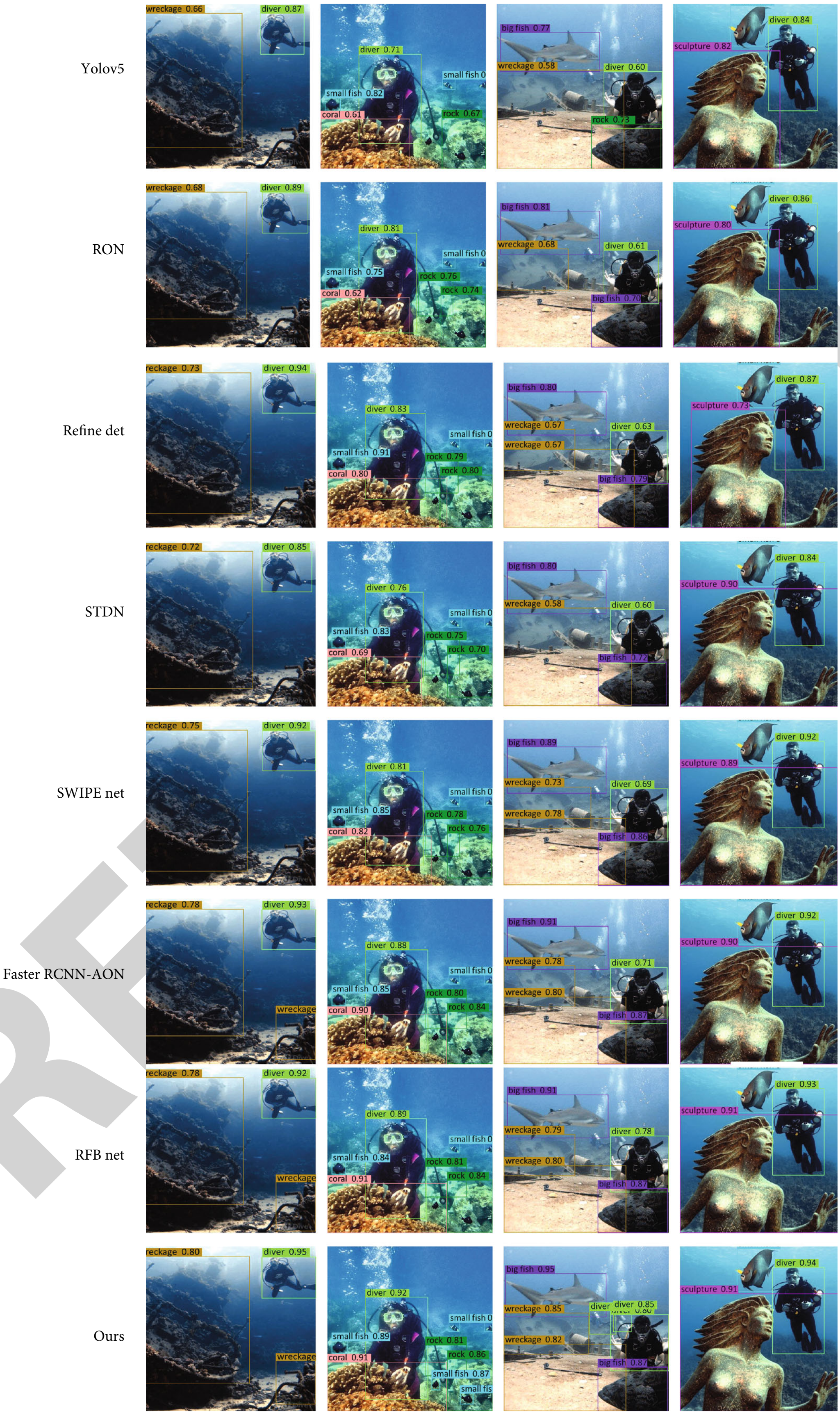


Figure 10: Visualization of the detection results on the CUID dataset compared with the state-of-the-art methods.

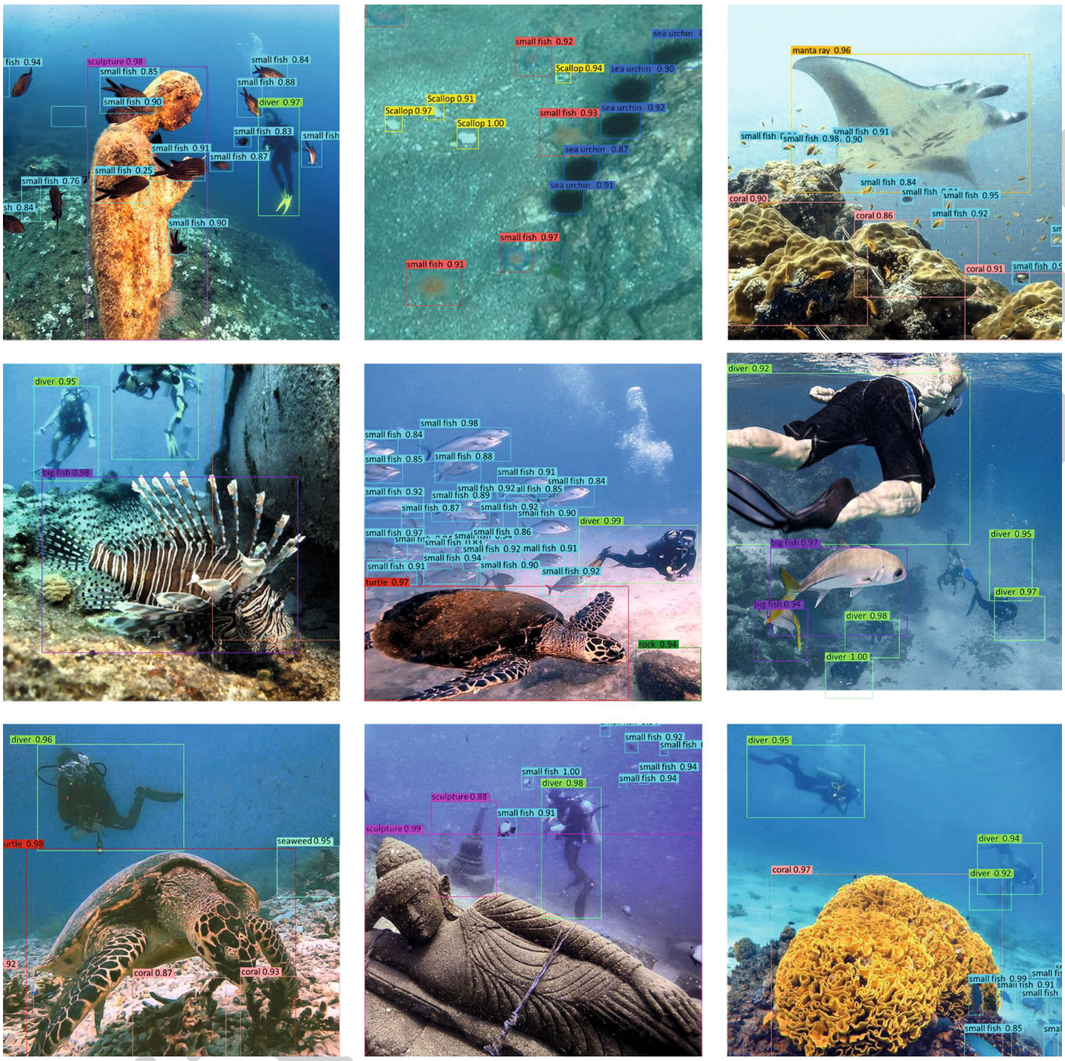


Figure 11: Visualization of our proposed method on the CUID dataset.

Table 5: Comparison of detection performance of embedded different modules. The best results are marked in bold.

Baseline	LFEM	CAM	SNFP	AP_S	AP_M	AP_L	AP_{50}	AP_{75}	mAP
✓				40.05	68.30	69.28	59.27	39.58	53.13
✓	✓			41.13	70.34	72.37	60.91	41.24	55.90
✓	✓		✓	44.95	72.65	77.74	65.32	46.82	58.68
✓	✓	✓		42.84	71.18	75.42	63.71	44.33	57.10
✓	✓	✓	✓	**48.73**	**76.90**	**83.41**	**69.84**	**49.94**	**60.12**

The mAP among 15 categories of objects is 81.94%, which is better than the state-of-the-art methods. From the perspective of a single category of targets, coral, rock, and sculpture have the worst detection effect. The main reason is that corals cannot be clearly distinguished from rocks, resulting in many false detections. The reason for the poor detection effect of sculptures is that some humanoid sculptures are classified as divers.

To further compare the detection effects of objects of different scales, Table 3 lists the performance of our method on the CUID dataset using the COCO indicator relative to Yolov5, RON, RefineDet, STDN, SWIPENet, Faster R-CNN-AON, and RFBNet. It can be seen from this table that the average detection accuracy of our method for small objects and large objects is the best, reaching 48.73% and 83.41%, respectively, which shows that the proposed method can well adapt to the scenario of multisize underwater objects and can accurately detect underwater objects of different scales. Meanwhile, our method also achieves the best detection effect under a stricter IOU, reaching 69.84% and 49.94% for AP50 and AP75, respectively, which can provide more accurate bounding boxes for multiscale objects.

In terms of detection speed, Table 4 shows the comparison results of parameters, model size, FLOPs, and FPS on Yolov5, RON, RefineDet, STDN, SWIPENet, Faster R-CNN-AON, RFBNet, and our proposed method on the CUID dataset. We can find that our method has fewer parameters, a smaller model size, and less computational resource consumption. Also, it has a relatively fast detection speed.

Figure 9 shows the *P-R* curves of detecting small, medium, and large objects and all objects on the CUID dataset. Obviously, our method can achieve the best results in detecting objects of various scales. In particular, when the recall rate is 0.5 to 0.7, the *P-R* curve of our method for detecting small objects is much higher than that of other detection networks. This indicates that when our method detects multiscale objects in underwater images with low image quality, the detection effect on small-scale objects is the most improved compared to other advanced methods. Overall, as a lightweight target detection network, our method can detect underwater multiscale targets quickly and effectively, and it achieves a good balance between detection accuracy and speed.

4.4.2. Subjective Evaluation. The visualization results of object detection on the CUID dataset are shown in Figures 10 and 11. It can be seen from Figure 10 that compared with other advanced methods, our proposed method can effectively reduce the missed detection rate, especially for small-scale objects, such as the small fishes in the second picture and the divers in the third picture. Figure 11 shows the detection results of our method on some other images of the dataset. Our proposed method can successfully detect objects of different scales. This is because our method achieves scale-aware contextual information aggregation and reduces the loss of effective information at low resolutions, while SNFP achieves adaptive and accurate detection of objects of different scales.

4.5. Ablation Experiment. To prove the rationality of the three functional modules proposed in our method, an ablation experiment was conducted to verify the effect of each module on object detection performance. Table 5 presents the ablation results of adding each module (namely, LFEM, CAM, and SNFP) to the MobileNetv3 framework. It can be seen that the adding of each module brings benefits to the whole network, especially the SNFP module, which can be seen from the comparison between model 3 and model 4. The network detection performance of model 5 is the highest, which indicates that the three modules are indispensable, and the combination of them leads to the best detection effect on underwater multiscale objects.

5. Conclusion and Future Work

This study proposes a lightweight underwater image object detection method. In the proposed method, MobileNetv3 is the backbone network for preliminary feature extraction. LFEM pays attention to the feature map at the channel and space levels. The features with heavy weights are promoted, and the features with small weights are suppressed. Meanwhile, the cross-group information exchange enriches the semantic information and location information of the objects. CIAM pools the extracted feature maps to obtain feature pyramids, and it fuses feature maps of different scales using the original URC method to realize an effective fusion of global context information and enhance the feature representation ability. The SNFP performs adaptive multiscale perception and multianchor detection on feature maps of different sizes to cover objects of different sizes and realize multiscale object detection in underwater images. Our proposed method can realize light feature extraction and effectively handle the global relationship between the scene and the object while expanding the receptive field, thus achieving adaptive multianchor detection on multiscale objects in underwater images.

The experimental results show that the average detection accuracy of our proposed method reaches 81.94, the model size is only 31.2 Mb, and the detection speed reaches 44.3 FPS. Overall, our proposed method outperforms the state-of-the-art methods in terms of detection accuracy, lightweight, and real-time performance. The proposed method can be used for effective underwater image multiscale object detection.

In future work, moving the proposed method to more application scenarios is the focus of our research. Also, the integration of image acquisition and detection using underwater intelligent robots will be explored.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest to report regarding the present study.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (Grant No. 61671470) and the Key Research and Development Program of China (Grant No. 2016YFC0802900).

References

[1] Q. Jiang, Y. Chen, and G. Wang, “A novel deep neural network for noise removal from underwater image,” *Signal Processing*, vol. 87, article 115921, 2020.

[2] M. Jaderberg, K. Simonyan, and A. Zisserman, “Spatial transformer networks,” *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[3] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, Salt Lake City, UT, 2018.

[4] X. Zhang, X. Zhou, M. Lin, and J. Sun, “Shufflenet: an extremely efficient convolutional neural network for mobile

devices," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018.

[5] M. Zhang, C. Liu, S. Wang, Q. He, and Q. Wei, "Lightweight underwater object detection based on YOLO v4 and multi-scale attentional feature fusion," *Remote Sensing*, vol. 13, no. 22, p. 4706, 2021.

[6] G. Ross, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014.

[7] R. Girshick, "Fast R-CNN," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Translation Pattern Analysis*, vol. 39, no. 6, pp. 1137–1149, 2017.

[9] W. Liu, D. Anguelov, D. Erhan et al., "SSD: single shot multibox detector," in *Computer Vision – ECCV 2016. ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9905 of Lecture Notes in Computer Science, Springer, Cham, 2016.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016.

[11] J. Redmon and F. Ali, "YOLO9000: better, faster, stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017.

[12] W. Fang, L. Wang, and P. Ren, "Tinier-YOLO: a real-time object detection method for constrained environments," *IEEE Access*, vol. 8, pp. 1935–1944, 2020.

[13] R. Huang, J. Pedoeem, and C. Chen, "YOLO-LITE: a real-time object detection algorithm optimized for non-GPU computers," in *2018 IEEE International Conference on Big Data (Big Data)*, Seattle, WA, USA, 2018.

[14] T. Kong, F. Sun, A. Yao, H. Liu, M. Lu, and Y. Chen, "Ron: reverse connection with objectness prior networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Honolulu, HI, USA, 2017.

[15] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018.

[16] T. Lin, P. Dollar, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017.

[17] M. Everingham, J. Laaksonen, and V. Viitaniemi, "The Pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[18] Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017.

[19] P. Zhou and G. Cong, "Scale-transferrable object detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018.

[20] L. Chen, Z. Liu, and L. Tong, "Underwater object detection using invert multi-class Adaboost with deep learning," in *2020 International Joint Conference on Neural Networks (IJCNN)*, Glasgow, UK, 2020.

[21] W. H. Lin, J. X. Zhong, S. Liu, T. Li, and G. Li, "RoIMix: proposal-fusion among multiple images for underwater object detection," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2019.

[22] P. Sahu, N. Gupta, and N. Sharma, "A survey on underwater image enhancement techniques," *International Journal of Computer Applications*, vol. 87, no. 13, pp. 19–23, 2014.

[23] L. Zeng, B. Sun, and D. Zhu, "Underwater target detection based on Faster R-CNN and adversarial occlusion network," *Engineering Applications of Artificial Intelligencevol*, vol. 100, no. 4, article 104190, 2021.

[24] Z. Cheng, Y. Xu, F. Bai, Y. Niu, S. Pu, and S. Zhou, "AON: towards arbitrarily-oriented text recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018.

[25] D. Haase and M. Amthor, "Rethinking depthwise separable convolutions: how intra-kernel correlations lead to improved mobilenets," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020.

[26] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018.

[27] A. Howard, M. Sandler, B. Chen et al., "Searching for mobilenetv3," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*,, Seoul, Korea (South), 2019.

[28] N. Ma, X. Zhang, H. T. Zheng, and J. Sun, "Shufflenet v2: practical guidelines for efficient cnn architecture design," in *Computer Vision – ECCV 2018. ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11218 of Lecture Notes in Computer Science, Springer, Cham, 2018.

[29] K. Han and Y. Wang, "Ghostnet: more features from cheap operations," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020.

[30] B. Singh and S. D. Larry, "An analysis of scale invariance in object detection - snip," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018.

[31] B. Singh, M. Najibi, and L. S. Davis, "Sniper: efficient multi-scale training," *Advances in neural information processing systems*, vol. 31, 2018.

[32] S. Liu, Q. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[34] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*, vol. 2, pp. 2169–2178, New York, NY, USA, 2016.

[35] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[36] S. Liu and H. Di, "Receptive field block net for accurate and fast object detection," in *Computer Vision – ECCV 2018. ECCV*

2018, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11215 of Lecture Notes in Computer Science, pp. 404–419, 2018.

[37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, Las Vegas, NV, USA, 2016.

[38] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Computer Vision – ECCV 2018. ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11211 of Lecture Notes in Computer Science, Springer, Cham, 2018.

[39] H. Li, P. Xiong, H. Fan, and J. Sun, "Dfanet: deep feature aggregation for real-time semantic segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019.

[40] J. He, Z. Deng, L. Zhou, Y. Wang, and Y. Qiao, "Adaptive pyramid context network for semantic segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019.

[41] M. Fan, S. Lai, J. Huang et al., "Rethinking bisenet for real-time semantic segmentation," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 2021.

[42] R. Liu, X. Fan, M. Zhu, M. Hou, and Z. Luo, "Real-world underwater enhancement: challenges, benchmarks, and solutions under natural light," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4861–4875, 2020.

[43] *Northeast Fisheries Science Center*, Habitat mapping camera(-Habcam), 2012.

[44] C. Li, C. Guo, W. Ren et al., "An underwater image enhancement benchmark dataset and beyond," *IEEE Translation Image Process*, vol. 29, pp. 4376–4389, 2019.