The Visual Computer (2024) 40:1299–1310
https://doi.org/10.1007/s00371-023-02849-3

ORIGINAL ARTICLE

Multi-scale aggregation feature pyramid with cornerness for underwater object detection

Xinbin Li[1] · Haifeng Yu[2] · Haiyang Chen[1]

Accepted: 9 March 2023 / Published online: 9 April 2023

Abstract

Underwater object detection is a fascinating but challengeable subject in computer vision. Features are difficult to extract due to the color cast and blur of underwater images. Moreover, given the small scale of the underwater object, some details will be lost after several layers of convolution. Therefore, a multi-scale aggregation feature pyramid network is proposed to integrate multi-scale features and improve underwater object detection performance. Specifically, a lightweight and efficient network is used to extract the basic features. A special subnet is designed to improve the feature extraction capability of the backbone network to enrich the detailed features of small underwater objects. In addition, a multi-scale feature pyramid is proposed to enrich feature map. Each feature map enhances contextual information through a combination of up-sampling and down-sampling. The centerness strategy of the fully convolutional one-stage object detection head is improved by adding corner point regression to enhance the recall rate of small objects. Generalized intersection over union (GIoU) instead of IoU can better reflect the degree of coincidence between the actual box and the predicted box. Therefore, the regression loss is changed to GIoU loss. This paper evaluates the network on the underwater image dataset and obtains 78.90% mAP. Meanwhile, the experiment on the PASCAL VOC datasets is conducted and gets 84.3% mAP.

Keywords Underwater object detection · Feature pyramid network · Generalized intersection over union · Centerness strategy

1 Introduction

Underwater object detection is an important and difficult subject in computer vision. The underwater object detection task has attracted people's attention gradually. In recent years, many popular networks based on deep learning have achieved good results on common datasets. However, the image quality captured by the camera is poor [1, 2] because of the underwater lighting conditions and environment. These methods are not ideal when applied to the underwater object detection task directly. Underwater images have problems such as low contrast, color bias and uneven illumination because of the scattering and attenuation of light transport in the water [3]. As a result, underwater image features are difficult to extract. Objects in underwater images are usually small in size because of the long distance from the camera and the small actual size of the objects [4]. It is necessary to design an accurate object detection network for the above problems.

At present, convolutional neural networks (CNNs) are the backbone of most models based on deep learning. Different convolution layers allow one to extract the characteristics of different scales for CNNs [5, 6]. In general, a high-level feature map provides rich semantic information and it is advantageous to the detection of large objects. The low-level characteristics have rich texture information, more conducive to small object detection [7]. Detail information is crucial for small object identification. It is very important to construct multi-scale features for more complex underwater object detection tasks, which include not only abundant texture features but also strong semantic features [8].

CNN can learn advanced semantic features and use single-scale input features for recognition. SSD [9] uses the hierarchical feature of CNN pyramid, the multi-scale feature map from multiple layers calculated by the forward process. Feature pyramid network (FPN) [10] aims to use the pyramid

✉ Haifeng Yu
yhf5170@163.com

[1] Institute of Electrical Engineering, Yanshan University, Qinhuangdao 066004, Hebei Province, China

[2] College of Electrical Engineering, North China University of Science and Technology, Tangshan 063210, China

form of CNN hierarchical features naturally to generate feature pyramids with strong semantic information at all scales.

A multi-scale feature pyramid architecture based on FPN is proposed to detect underwater objects. Firstly, improved VoVNet [11] is taken as the backbone network. The one-shot aggregation (OSA) module in VoVNet only aggregates all the layers before the last one-time aggregation, which is highly efficient on the GPU and has fewer layers than the residual network layer at the same level. It can keep more details that are crucial to feature maps extraction. Secondly, a multi-scale aggregation feature pyramid is built. The basic features extracted from the backbone network are downsampled to enhanced features scale. Only top-down and horizontal connections result in the main information still come from the top. A new bottom-up network structure added to get a new feature pyramid. Then, the distance between corner point and bounding box is introduced to improve the recall of small objects. At the same time, the scale of the feature map is divided by the corresponding step size to better adapt to the size of the FPN. Batch normalization (BN) is replaced by group normalization (GN) after the convolutional layer of the head. GN divides the characteristic channels into groups, calculates the mean and variance to normalize within each group so that its calculation normalization will not depend on the batch size. Finally, GIoU [12] is added to measure the distance between the real box and the prediction box without overlap. GIoU pays attention to overlapping areas as well as other non-overlapping areas, which can better reflect the degree of overlap. The GIoU loss is added to the regression loss to ensure the accuracy of the prediction box. To summarize, the major contributions of this work are three-fold:

1. A convolution block of backbone is designed to enhance detailed information. The stage 1 of the original VoVNet-39 is replaced with two-channel convolution block. The added convolution block is used to extract the rich detail features of the image which are more conducive to the detection of small scale objects.
2. The feature scale is extended by downsampling based on FPN. An aggregation path from low-level features to high-level features is added to extract details. Therefore, a multi-scale aggregation feature pyramid is constructed. This structure can use context information to strengthen features and enhance the resolution of feature map.
3. The cornerness strategy is designed to add recall points. The distance between corner points and bounding boxes is introduced to add regression point and divided by the corresponding stride to improve the recall rate of small objects. The cornerness loss is designed based on the above method. Besides, IoU is replaced by GIoU as a measure of the actual box and the predicted box without overlap.

The rest of the paper is organized as follows. Section 2 presents related work about the development of technologies involved in our method. Section 3 describes the proposed methods specifically. Section 4 gives the experiments and analysis with proposed methods. Moreover, the last section presents conclusions on this work.

2 Related works

2.1 Object detection

Object detection is a heavily researched topic in computer vision. There has been a large body of researches on object detection with deep learning. According to whether region proposal is needed, popular object detection methods based on CNN mainly include two-stage object detection network [9, 30] and one-stage object detection network [15, 16].

The two-stage object detection network first extracts the candidate box from the image, then makes two corrections based on the candidate area to get the detection result. One-stage object detection networks remove region proposals unlike two-stage methods and directly regresses the location and category of the object. The latter can bring to faster detection. However, these methods all require presetting dense anchor which will introduces a lot of hyperparameters and is time consuming. The detection result is greatly affected by hyperparameters. Therefore, an anchor-free object detection model is constructed based on FCOS [22].

The backbone is responsible for extracting basic features from images in the object detection model, which is very important for object location and classification. ResNet [28] is the most commonly used backbone of object detection model. In fact, DenseNet [29] has stronger feature extraction ability than ResNet. Although it has a good effect for object detection with slow speed. The high memory access costs and power consumption are caused by dense connections in DenseNet. VoVNet [11] is designed to solve this problem. The object detection model based on VoVNet outperforms the model based on DenseNet with faster speed and better performance. VoVnet is selected as the backbone and extracts basic features as the input of FPN.

2.2 Multi-scale features

Recently, extracting features from different layers is popular in image recognition and these features are used together to detect objects. SPP-Net [13], R-CNN [14], Fast R-CNN [15] and Faster R-CNN [16] just take the final feature maps to detect object. Shrivastava et al [17] and SNIP [18] adopted the feature image pyramid, input images of different scales as image pyramids to generate features of different scales for prediction. The high accuracy is achieved at a high cost

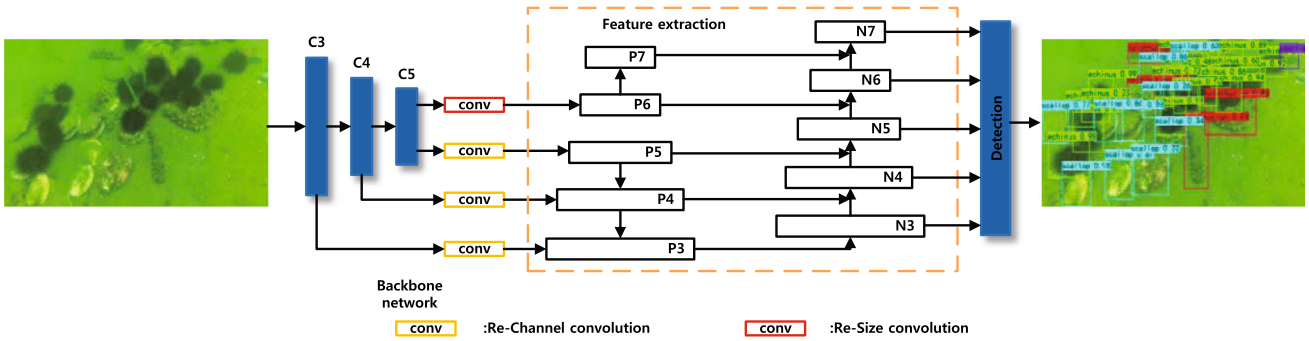


Fig. 1 The whole architecture of proposed network

in terms of time and memory. SSD, DSSD [19], YOLOv3 [20] are detecting objects in the feature pyramid extracted from inherent layers within the network while merely taking a single-scale image. This strategy ignores the context information of features. FPN utilized lateral connections and a top-down pathway to produce a feature pyramid and achieve more powerful representations. FCOS [22] and Xu et al [23] adjust the feature maps on FPN and take higher level feature maps to predict object. The top-down path results in information only coming from the top and information from the bottom is not well utilized. M2Det [21] uses U-shaped pyramid to extract feature depth, concat features of different levels according to the same scale. Nas-FPN [24] uses neural network search based on RetinaNet [25] to design FPN. However, these improved FPN models are complicated and the search cost is high. Therefore, a multi-scale aggregation feature pyramid is constructed based on FPN.

2.3 FCOS detection head

The detection head detects the location and category of objects based on the features obtained from above. Detection heads can be divided into anchor-based and anchor-free models. Anchor-based model need predefined anchor point generation candidate boxes such as YOLO and SSD. The anchor involves many hyperparameters which have a great influence on the final result. The anchor-free model is based on corner or center such as CenterNet [31] and FCOS [22]. The FCOS network adopts the regression strategy of anchor-free. Although the recall rate is improved, it will produce many prediction bounding boxes with low quality center point offset. Based on this, a simple and effective strategy centerness is proposed to suppress these low quality bounding boxes. This strategy does not introduce any hyperparameters. The FCOS detection head is selected as the basic detection head. Because it is a general object detection network, this paper will improve this detection head for underwater environment.

The loss function is used to estimate the difference between the predicted value and the real value of the model. Intersection-over-union (IoU) loss is used as regression loss in FCOS. IoU calculates the ratio of the intersection and union between the predicted box and the actual box. The IoU loss does not provide any motion gradients and cannot be trained when the predicted box and the actual box do not overlap. Chen [27] proposes Pixel-IoU loss to improve the accuracy of both rotation angle and IoU. Zheng [26] takes into account the distance, overlap rate and scale to design distance-IoU (DIoU), making the object box regression more stable. CIoU is further proposed on the basis of DIoU, the length-width ratio of the three elements of box regression is considered in the calculation. However, the centers of dense objects in underwater images are close. They will be removed after non-maximum suppression (NMS) processing. Hamid put forward the idea of generalized intersection over union, introducing the smallest bounding rectangle of the predicted box and the actual box on the basis of the IoU. The predicted boxes will move towards the object box given the introduction of penalty terms. It overcomes the above shortcomings of the IoU.

3 Proposed method

This paper proposes path aggregation feature pyramid network to settle the issue on underwater object detection. The model architecture is represented in Fig. 1. Our network will be introduced from three aspects: feature extraction, multi-scale feature fusion and object detection.

3.1 Feature extraction

This paper proposes feature extraction architecture based on VoVNet to obtain abundant and robust feature maps. A refined convolutional block is proposed to replace the backbone of the first convolutional layer. For VoVNet-39, the first convolutional layer uses multi-layer convolution to

Table 1 Two constructs of VoVNets Stage1

Type	Output stride	VoVNet-39-A	VoVNet-39-B
		[3×3conv 64 stride=2	
Stem	2	3×3conv 64 stride=1	
Stage 1	2	3×3conv 128 stride=1]	3×3conv 128 stride=2
	2	&	
		[3×3conv 128 stride=2]	

extract the feature, which loses more detail information of small objects than single-layer convolution. Dual convolution block is used to extract image details. The two channels can extract more abundant basic feature, so that they are more conducive to small object detection. Then, the improved VoVNet-39 as the backbone is called VoVNet-39-A. Three OSA modules are used to aggregate all the previous layers once after a refined convolutional layer. The output of each OSA module is used as the basic feature to generate the basic feature layer at different scales. A branch is added to VoVNet stage1 to extract richer details. The added branches are used alone to compare and verify the effect of feature extraction. The structure of stage 1 is shown in Table 1.

Finally, a multi-scale feature pyramid is constructed. The basic features are obtained from the OSA module output of the backbone network, while the high level features are sampled from the basic feature map. A new feature pyramid is obtained by adding a new bottom-up network structure. This pyramid feature map integrates feature maps of different sizes from low level to high level, contains rich texture information and semantic information which are beneficial to underwater object detection.

3.2 Multi-scale feature pyramid

This paper builds a multi-scale feature pyramid to acquire robust feature maps. Inspired by FPN, the third to fifth convolutional blocks are taken to extract feature maps and build our deeper feature pyramid. We conduct upsampling from higher level feature maps to enhance lower level feature map with context information. Our feature pyramids are combined with five feature maps, where each feature map has a different scale. This paper builds a multi-scale feature pyramid based on feature extraction network. Our multi-scale feature maps are defined as P_3, P_4, P_5, P_6, P_7 and N_3, N_4, N_5, N_6, N_7, where the strides of them are 8, 16, 32, 64, 128, respectively. C_3, C_4, and C_5 are the initial feature layers and the scaling process can be described as:

$$\begin{aligned} P_{\mathrm{i}} &= f_1 * C_{\mathrm{i}} + \mu * P_{\mathrm{i}+1} \quad i = 3, 4 \\ P_5 &= f_1 * C_5 \\ P_6 &= f_2 * C_5 \\ P_7 &= f_2 * P_6 \end{aligned} \tag{1}$$

P_{i} represents the i-th layer feature of the P level feature pyramid, f_1 is the variable channel number filter with the convolution kernel of 3×3 and the stride of 1, f_2 is the down-sampling filter with the convolution kernel of 3×3 and the stride of 2, μ is upsamping,* is the convolution operation.

Each building block takes a higher resolution feature map N_{i} and a coarser map $P_{\mathrm{i}+1}$ through lateral connection and generates the new feature map $N_{\mathrm{i}+1}$. Each feature map N_{i} first goes through a 3×3 convolutional layer with stride 2 to reduce the spatial size. Then each element of feature map $P_{\mathrm{i}+1}$ and the down-sampled map are added through lateral connection. The fused feature map is then processed by another 3×3 convolutional layer to generate $N_{\mathrm{i}+1}$ for following sub-networks. This is an iterative process and terminates after approaching P_7. The 256-channel feature map is used in these features to be detected.

The feature fusion process can be formulated as follows,

$$N_{\mathrm{i}+1} = f_2 * N_{\mathrm{i}} + P_{\mathrm{i}+1} \quad i = 3...6 \tag{2}$$

N_{i} is the ith layer feature of the feature pyramid.

3.3 Detection of head

The FCOS head is selected as the base head and GN is added after the convolutional layer of the head so that its calculation during Normalization will not depend on the batch size value. The error of training and verification is higher when the batch size value is small. The calculation accuracy of the BN layer depends on the value of the current batch. GN divides the channels into groups and calculates the mean value and variance within each group for normalization. So it is not constrained by batchsize naturally.

Centerness as a unique branch of FCOS, the image is divided into grids according to scale. The training target is the distance between the center point of the grid and the truth value box. (x_0, y_0) and (x_1, y_1) are the corner points of the truth value box and (x, y) is the center points of the grid. The distances from the center point to the truth value box are, respectively, l^*, t^*, r^*, b^* :

$$\begin{aligned} l^* &= x - x_0 \quad t^* = y - y_0 \\ r^* &= x_1 - x \quad b^* = y_1 - y \end{aligned} \tag{3}$$

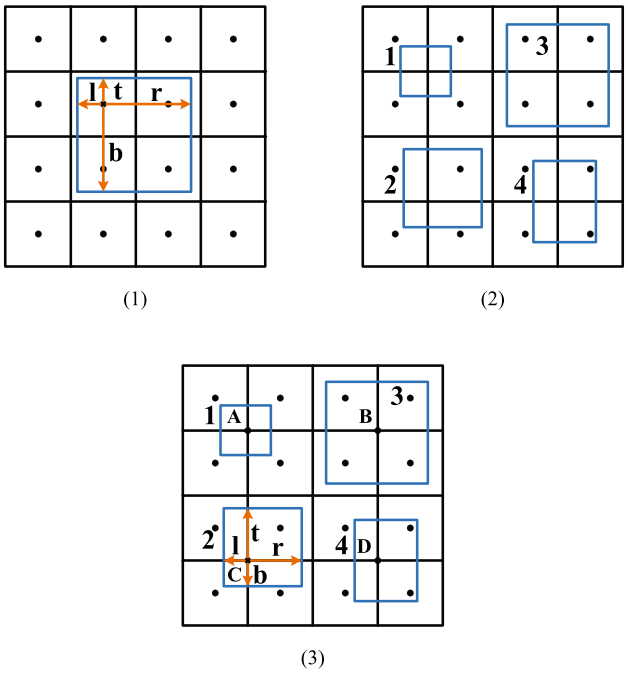


Fig. 2 The cornerness strategy

Centerness can be expressed as:

$$\text{Centerness}^* = \sqrt{\frac{\min(l^*, r^*)}{\max(l^*, r^*)} \times \frac{\min(t^*, b^*)}{\max(t^*, b^*)}} \tag{4}$$

However, there may be no center point or only one center point falling into the truth box when the object size is small. The scale of underwater image object is small, so the method needs to be improved in underwater object detection. We add corner points into the regression strategy to solve this problem. Corner points regression strategy is shown in Fig. 2.

In practical applications, the distance between the corner point and the actual box is divided by the corresponding stride to match the actual size of the underwater object. Then the distance between the corner point and the truth value box is, respectively, $l_c^*, t_c^*, r_c^*, b_c^*$:

$$\begin{aligned} l_c^* &= (x - x_0)/s \quad t_c^* = (y - y_0)/s \\ r_c^* &= (x_1 - x)/s \quad b_c^* = (y_1 - y)/s \end{aligned} \tag{5}$$

The Cornerness is:

$$\begin{aligned} l_c^* &= (x - x_0)/s \quad t_c^* = (y - y_0)/s \\ r_c^* &= (x_1 - x)/s \quad b_c^* = (y_1 - y)/s \end{aligned} \tag{6}$$

The loss function of FCOS network is

$$L_1 = \frac{1}{N_{\text{pos}}} \sum\nolimits_{x,y} L_{\text{cls}}(c_{\text{x,y}}, c_{\text{x,y}}^*) + \frac{\lambda}{N_{\text{pos}}} \sum\nolimits_{x,y} 1_{c_{\text{x,y}}^* > 0} \times L_{\text{reg}}(t_{\text{x,y}}, t_{\text{x,y}}^*) + \frac{1}{N_{\text{pos}}} \sum\nolimits_{x,y} L_{\text{cen}}(e_{\text{x,y}}, e_{\text{x,y}}^*) \tag{7}$$

where L_{cls} is focal loss, L_{reg} is the IoU loss and L_{cen} is Centerness loss. N_{pos} denotes the number of positive samples and λ is 1 in this paper is the balance weight for L_{reg}. $c_{\text{x,y}}^*$ is the true values of the target category, $t_{\text{x,y}}^*$ is the true values of the target position, $e_{\text{x,y}}^*$ is the true values of Centerness, the predicted target category is $c_{\text{x,y}}$, the predicted target position is $t_{\text{x,y}}$ and the predicted Centerness is $e_{\text{x,y}}$. $1_{c_{\text{x,y}}^* > 0}$ is the activation function, being 1 if $c_{\text{x,y}}^* > 0$ and 0 otherwise.

The regression process is improved as well as the corresponding loss function. The new loss function is added with corner regression to form a new loss function:

$$L_2 = \frac{1}{N_{\text{pos}}} \sum\nolimits_{x,y} L_{\text{cls}}(c_{\text{x,y}}, c_{\text{x,y}}^*) + \frac{\lambda}{N_{\text{pos}}} \sum\nolimits_{x,y} 1_{c_{\text{x,y}}^* > 0} \times L_{\text{reg}}(t_{\text{x,y}}, t_{\text{x,y}}^*) + \frac{1}{N_{\text{pos}}} \sum\nolimits_{x,y} L_{\text{cor}}(e_{\text{x,y}}, e_{\text{x,y}}^*) \tag{8}$$

where L_{cor} is Cornerness loss. IoU calculates the ratio of the intersection and union of the predicted box and the actual box. However, The IoU has the disadvantage of not measuring the distance between two boxes and the way of intersection. GIoU aims to overcome the shortcomings of IoU and takes full advantage of it. IoU can be propagated back for intersecting boxes, it can be directly used as the objective function of optimization. But the gradient will be zero and optimization cannot be performed if they do not intersect. Using GIoU at this point completely avoids this problem. The regression loss is replaced by GIoU loss to form an objective function. The training loss function is defined as the sum of L_1 and L_2.

4 Experiments and analysis

In this section, we design several group experiments of proposed method and analysis of results to verify our work. Our experiments are mainly conducted on 4 categories of an underwater image dataset. The experiment section includes 4 parts: (1) introducing implementation details about the experiments; (2) experiments on underwater image datasets; (3) analysis of the loss function; (4) experiments on PASCAL VOC datasets.

4.1 Implementation details

We implement MA-FPN and other networks based on PyTorch. The VoVNet-39-A is taken as our backbone networks. Specifically, our network is trained with stochastic gradient descent (SGD) for 100K iterations with the initial

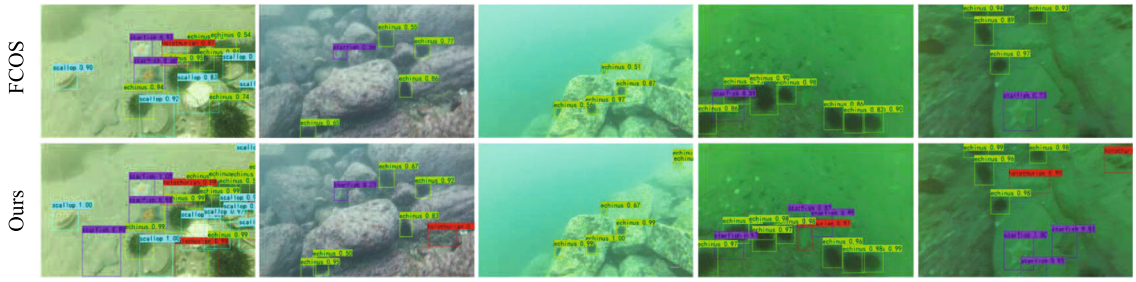


Fig. 3 Comparison of detection results between baseline and our proposed method

learning rate being 0.01 and a minibatch of 4 images. The learning rate is reduced by a factor of 10 at iteration 60K and 80K, respectively. Weight decay and momentum are set as 0.0001 and 0.9, respectively. In addition, the input images are resized to 1024*800. All of the experimental results are implemented using a Nvidia TiTan Xp GPU and cuDNN v5.1 and an Intel(R) Xeon(R) W-2135 CPU@3.70GHz. The dataset is available at https://aistudio.baidu.com/aistudio/datasetdetail/25886 publicly.

4.2 Experiments on underwater image datasets

The underwater image datasets are built with the same format of PASCAL VOC datasets, which mainly include 5546 pictures with four categories: starfish, echinus, holothurian and scallop. The underwater images are blur and color cast. The scales of underwater objects in underwater images are small. What is more, some underwater objects have protective coloration to hide themselves into surroundings, such as holothurian and scallops.

The captured images usually have a high density of objects because of the living habits of underwater objects. These natures aggravate the challenges of underwater object detection task. A series of experiments are performed on underwater image dataset in accordance with the proposed network. The detection results of FCOS and the improved network are selected to verify the performance of the method. It can be seen from the comparison in Fig. 3 that our method is better than FCOS in underwater object detection. Specifically, the improved network can detect more small objects and objects with protective colors.

4.2.1 Comparison with popular detectors

Experiments were carried out with different detectors on the underwater image dataset. Specifically, each popular detector was reimplemented with default settings on the underwater image dataset. The comparison results are shown in Table 2. Obviously, the underwater object detection performance cannot reach the common type of detection performance.

Table 2 Comparison with popular detectors on the underwater image dataset

Approach	Backbone	Input size	FPS	mAP (%)
Faster-RCNN [16]	VGGNet	1000×600	7	71.18
	ResNet	1000×600	5.8	70.34
SSD512 [9]	VGGNet	512×512	9.1	72.51
YOLOv3 [20]	DarkNet	416×416	**16.8**	73.60
YOLOv4 [30]	DarkNet	608×608	14.5	76.46
FPN [10]	ResNet	800×1024	6.5	74.85
FCOS [22]	ResNet	800×1024	4.3	74.53
MA-FPN	VoVNet	800×1024	4	**78.90**

Bold values indicate the best results

In object detection tasks, SSD, YOLO and RCNN series are popular methods. This article implements these networks on the same underwater dataset. As shown in Table 2, the mAP of the two-stage object detection network Faster-RCNN on the underwater dataset is 71.18%. It has higher detection accuracy compared with the single-stage object detection network SSD and YOLOv3, but the complex network structure causes the detection speed to be low. The YOLOv3 detector achieves a faster detection speed and can process 16.8 frames per second. YOLOv4 [30] improved the detection accuracy of underwater dataset to 76.46%. The SSD detector obtains a detection accuracy of 72.51% through the feature pyramid. Our proposed method performs best on the underwater image dataset with a mAP of 78.90%. Below we will analyze the effectiveness of our network in detail.

4.2.2 Ablation study

We conduct a series of ablation experiments to show the comparative effect of each component for verifying performance of proposed network. In Table 3, the FCOS on underwater image dataset is considered as baseline and introduce our design on it to improve the performance.

The comparison between the second line and the third line shows that the underwater object detection performance is improved after the introduction of MA-FPN, which is attributed to the rich texture information and semantic infor-

Table 3 Ablation experiments on underwater image dataset

Network	AP (%)				mAP
	Starfish	Echinus	Holothurian	Scallop	(%)
Origin	86.24	86.98	58.99	65.89	74.53
VoVNet-39-A	86.50	89.45	61.64	68.97	76.64
VoVNet-39-A FPN	85.85	89.79	63.10	69.70	77.11
VoVNet-39-A FPN GIoU GN	85.92	89.85	66.52	69.72	78.00
VoVNet-39-A FPN GIoU GN Cornerness	**87.02**	**90.16**	**67.65**	**70.71**	**78.90**

Bold values indicate the best results

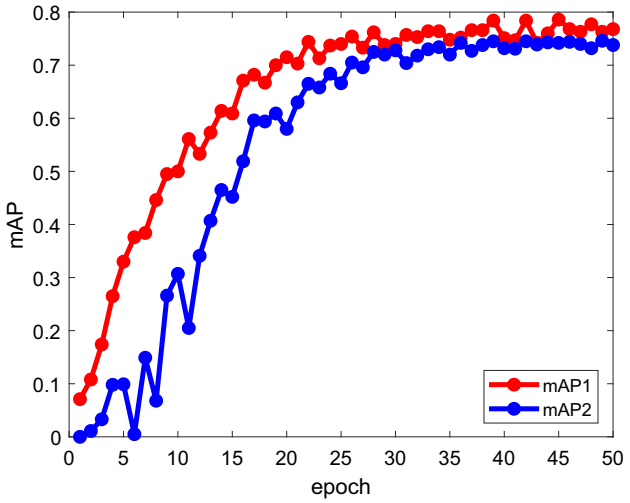


Fig. 4 Comparison of detection results of two networks. Blue line is the mAP of FCOS and red line is the mAP of MA-FPN

mation of MA-FPN. What is more, the results of the third and fourth lines show that the redesign of the loss function also contributes to the detection performance. This is because GIoU can better reflect the coincidence degree of the prediction box and ground truth. It can be seen from the last two lines that the corner point regression strategy designed by us is effective. It takes into account more regression points and has a more friend detection effect for small objects existing underwater. The proposed network has advantage on underwater object detection by contrast with FCOS. MA-FPN performance is 4.37% better than FCOS with the same setting of experiments.

Figure 4 shows the FCOS network and the underwater dataset detection results of the proposed network. First of all, at the beginning of training, our network has a higher mAP improvement effect than FCOS in the same epoch and the fitting speed is fast. On the other hand, the mAP of MA-FPN is higher than FCOS for each epoch, which proves the effectiveness of the proposed network.

4.2.3 Research on backbone

The experiments are carried out on different types of backbone networks to discuss the influence of different backbone networks on detection performance. The experimental results are shown in Table 4. The network using ResNet as the backbone network only generates 74.53% of mAP on the underwater image dataset. The overall detection effect of VoVNet is better than ResNet. VoVNet-39-A is more suitable for our network compared with the previous three lines.

4.3 Analysis of the loss function

After several simulation experiments, our network achieves good performance in underwater object detection. The comparison of precision-recall curve between FCOS and ours is shown in Fig. 5.

Neural network training is to reduce the loss function continuously. The fitting effect of the model can be judged by comparing the loss function when there is no change in the dataset. Fig. 6 is proved by experimental data. Loss1 and loss2 are the training loss reduction curves of the FCOS and proposed network. It can be found by comparison when the network training loss value is stable, the loss function

Table 4 Influence of backbone network structure on detection performance

Backbone	AP (%)				mAP
	Starfish	Echinus	Holothurian	Scallop	(%)
ResNet-50	86.24	86.98	58.99	65.89	74.53
VoVNet-39	86.61	87.83	60.05	**69.51**	76.00
VoVNet-39-B	**87.07**	87.79	59.51	68.59	75.74
VoVNet-39-A	86.50	**89.45**	**61.64**	68.97	**76.64**

Bold values indicate the best results

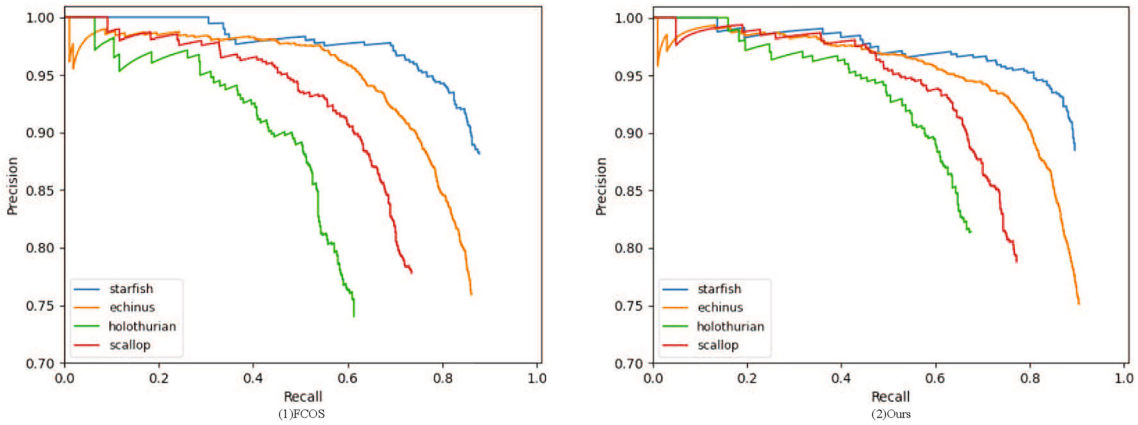


Fig. 5 The comparison of precision-recall curve between FCOS and Ours

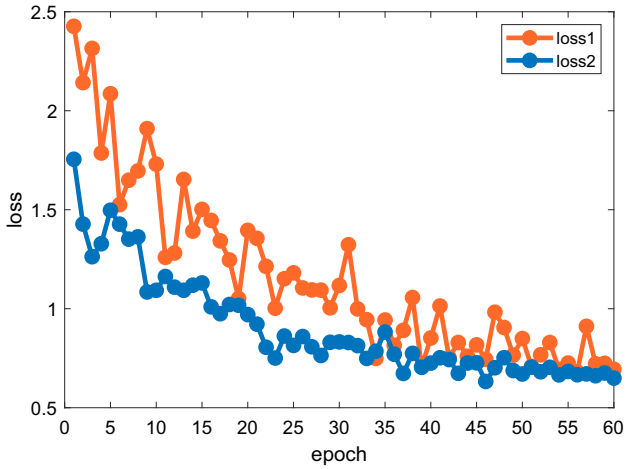


Fig. 6 The comparison of training loss

designed by us has smaller value and better data fitting. Moreover, it can be seen the network is more stable since the loss function of our network fluctuates less during the training process. Among them, the cornerness strategy plays a crucial role. As shown in Fig. 7, under the condition of numerous small-size objects in underwater images, the corner point regression strategy added can detect more small objects by setting more recall points.

More detection results of our network are shown in Fig. 8. It can be seen that our network performs well in scenes with fuzzy and uneven lighting. Even if there are a large number of small objects, the objects can be accurately detected.

However, this network still has shortcomings in underwater object detection. It is extremely difficult to identify occluded objects. In addition, the detection of covered objects also needs to be improved. Some failed detection cases are shown in Fig. 9. For example, it is difficult to detect a starfish hidden behind a stone. When a sea urchin is very close to a mesh with similar characteristics, it will be ignored as a mesh.

To verify the practicability of the network, we simulated a real underwater environment in a laboratory pool. As shown in Fig. 10, the underwater robot collects underwater images of the object to be detected through a camera, it uses our network to detect these objects. It can be seen that our network can accurately detect underwater objects under certain conditions and has certain practicability.

4.4 Experiments on PASCAL VOC datasets

We conducted experiments on the PASCAL VOC dataset to verify the effect of proposed method. Specifically, we train the network on the VOC 2007 and VOC 2012 training sets, then test the model on the VOC 2007 test set. We compare our network with the latest object detection networks on the PASCAL VOC dataset in Table 5.

As shown in Table 5, YOLOv3 can detect objects in real time at a speed of 34 frames per second and SSD300 can reach 46 FPS on detection tasks. The upgraded networks of these detectors, such as SSD512 and DSSD321, achieve higher detection accuracy at the cost of increasing the computational burden. DSSD321 even reached 78.6% of mAP. Our network obtained the highest mAP value 84.3% on the PASCAL VOC dataset, exceeded the FCOS by 3.8% mAP.

4.5 Robust testing experiments

In this section, we analyze the object detection accuracy of the proposed network in noisy environments to verify the robustness [32]. As shown in Fig. 11, we add Gaussian noise obeying a normal distribution $N\left(\mu, \sigma^2\right)$ to validate the proposed network. The abscissa of Fig. 11 is the noise parameter σ, and the ordinate is the average detection accuracy mAP.

The values of mAP are 0.769, 0.715, 0.554, 0.340, respectively, when σ is 0.1, 0.3, 0.5, 0.8. It can be seen that the proposed object detection method is robust to a certain extent.

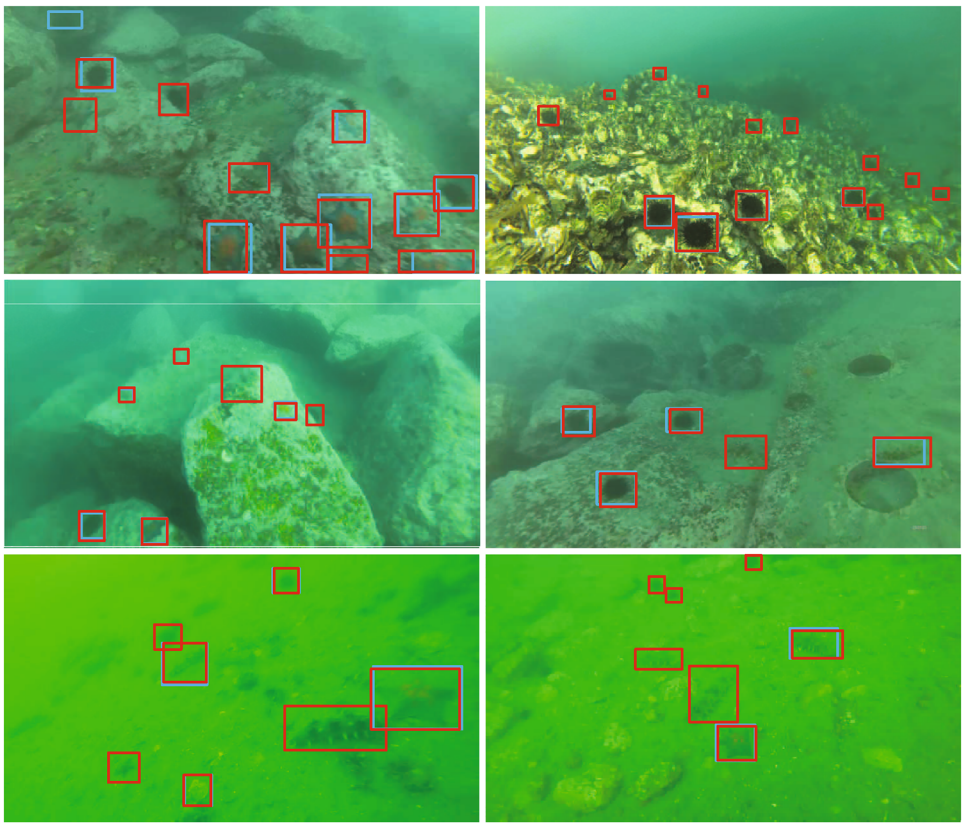


Fig. 7 Some detection results of FCOS and MA-FPN on underwater image dataset. Blue bounding boxes are the FCOS detection results and red ones represent the MA-FPN detection results

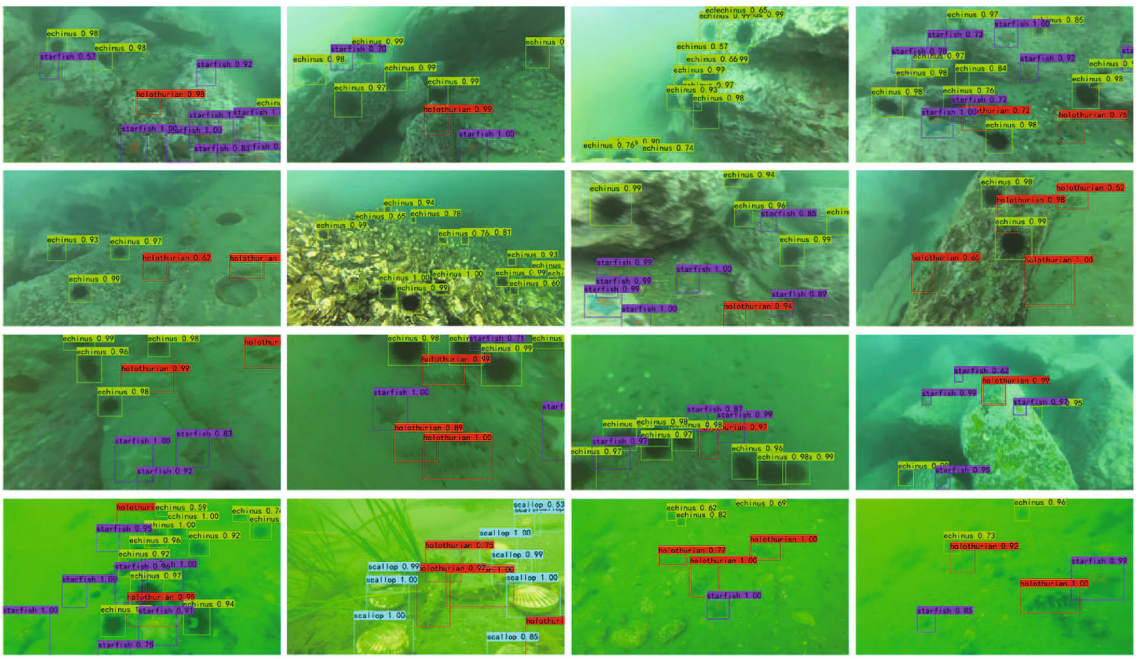


Fig. 8 Part of the detection results of our method on the underwater image dataset

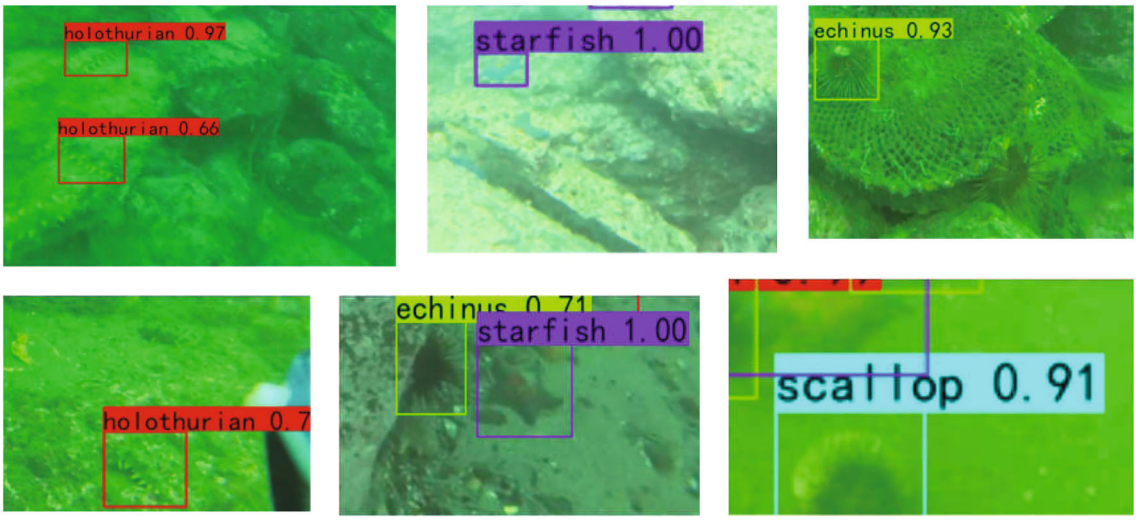


Fig. 9 Failure cases on underwater object detection task

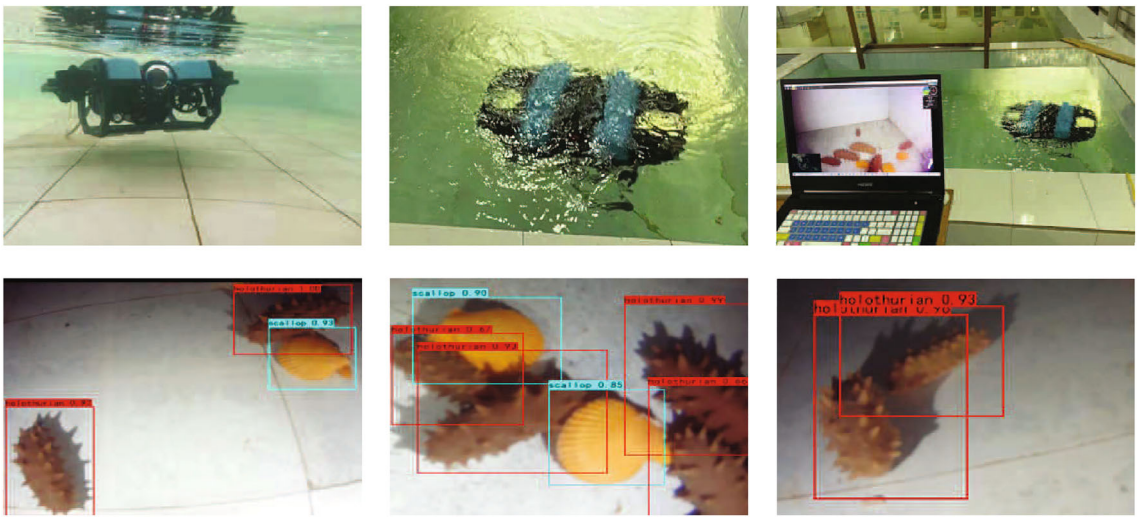


Fig. 10 Simulation of underwater environment detection results

Table 5 Detection results on the PASCAL VOC datasets

Approach	Backbone	Input size	FPS	mAP (%)
Fast-RCNN [15]	VGGNet	1000×600	0.6	70.0
Faster-RCNN [16]	VGGNet	1000×600	7	73.2
	ResNet	1000×600	5	76.4
FPN [10]	ResNet	1280×768	5	77.1
SSD300 [9]	VGGNet	300×300	**46**	74.3
SSD512 [9]	VGGNet	512×512	19	76.8
DSSD321 [19]	ResNet	321×321	9.5	78.6
YOLOv3 [20]	DarkNet	416×416	34	77.2
FCOS [22]	ResNet	1024×800	5.2	80.5
MA-FPN	VoVNet	1024×800	4.9	**84.3**

Bold values indicate the best results

Our network is very robust when the added noise is small. However, the underwater image itself has strong noise, so when a large amount of noise is added, the detection accuracy is affected to a certain extent.

5 Conclusion

This paper proposes a simple and effective multi-scale feature pyramid network structure, which is used to construct a feature pyramid to detect multi-scale objects. First, the efficient VovNet-39-A is selected as the backbone network to extract the basic features. Then, a multi-scale feature pyramid is built to enhance the texture and semantic features. In addition, the corner point regression strategy is introduced and divide it by the scale of the feature when calculating the point regression to adapt to the actual scale of the object. Finally, this paper uses GIoU instead of IoU to improve the loss function to measure the distance between the prediction

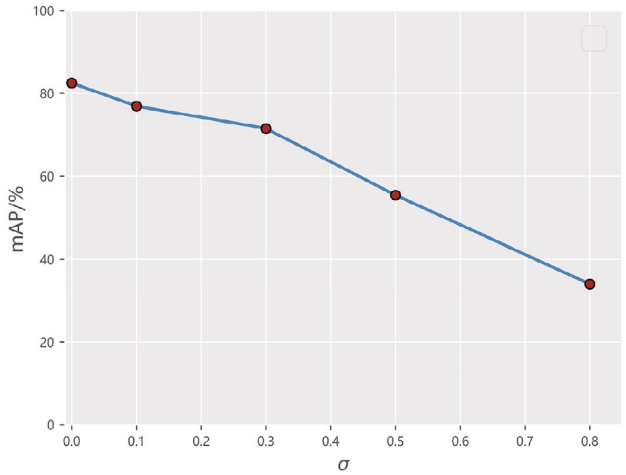


Fig. 11 Object detection accuracy under different degrees of Gaussian noise

box and the real box. Experimental results show that this method is effective for underwater object detection. After several experiments, the map of this method in underwater object detection reaches 78.90%, which is better than FCOS by 4.37%.

Acknowledgements This work was supported in part by the Hebei Natural Science Foundation, China under Grant F2020203037, and F2022203025, in part by the National Natural Science Foundation of China under Grant 61873224, Grant 62271437, and Grant 62003295, in part by the Science and Technology Research Project of Universities in Hebei, China under Grant QN2020301.

Data availability The data that support the findings of this study are available from Peng Cheng Laboratory. Restrictions apply to the availability of these data, which were used under license for this study. Data are available at https://aistudio.baidu.com/aistudio/datasetdetail/25886 with the permission of Peng Cheng Laboratory.

Declarations

Conflict of interest The authors declared that they have no conflicts of interest to this work.

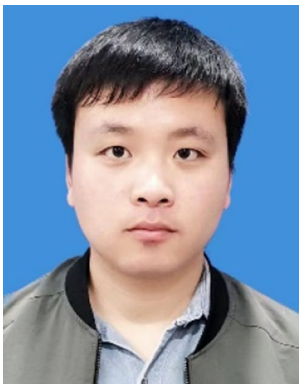
References

1. Han, M., et al.: A review on intelligence dehazing and color restoration for underwater images. IEEE Trans. Syst. Man Cybern. Syst. **50**(5), 1820–1832 (2018)
2. Wang, Jing, et al.: CA-GAN: class-condition attention GAN for underwater image enhancement. IEEE Access **8**, 130719–130728 (2020)
3. Wang, Xinhua, et al.: Underwater object recognition based on deep encoding-decoding network. J. Ocean Univ. China **18**(2), 376–382 (2019)
4. Chen, L., et al.: Underwater object detection using invert multi-class adaboost with deep learning. In: 2020 International Joint Conference on Neural Networks (IJCNN). IEEE (2020)
5. Wei, Jian, et al.: Enhanced object detection with deep convolutional neural networks for advanced driving assistance. IEEE Trans. Intell. Transp. Syst. **21**(4), 1572–1583 (2019)
6. Dhillon, Anamika, Verma, Gyanendra K.: Convolutional neural network: a review of models, methodologies and applications to object detection. Progress Artif. Intell. **9**(2), 85–112 (2020)
7. Li, H., et al.: Pyramid attention network for semantic segmentation. arXiv:1805.10180 (2018)
8. Ammari, Habib, et al.: Reconstructing fine details of small objects by using plasmonic spectroscopic data. SIAM J. Imag. Sci. **11**(1), 1–23 (2018)
9. Liu, W., et al.: Ssd: single shot multibox detector. In: European conference on computer vision. Springer, Cham (2016)
10. Lin, T.-Y., et al.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
11. Lee, Y., et al.: An energy and GPU-computation efficient backbone network for real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019)
12. Rezatofighi, H., et al.: Generalized intersection over union: a metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
13. He, Kaiming, et al.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1904–1916 (2015)
14. Girshick, R., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2014)
15. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision (2015)
16. Ren, S., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern. Anal. Mach. Intell. 39(6):1137–1149 (2017)
17. Shrivastava, A., Abhinav, G., Ross, G.: Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
18. Singh, B., Larry, S.D. An analysis of scale invariance in object detection snip. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
19. Fu, C.-Y., et al.: Dssd: deconvolutional single shot detector. arXiv:1701.06659 (2017)
20. Redmon, J., Ali, F.: Yolov3: an incremental improvement. arXiv:1804.02767 (2018)
21. Zhao, Q., et al.: M2det: a single-shot object detector based on multi-level feature pyramid network. In: Proceedings of the AAAI Conference on Artificial Intelligence vol. 33. No. 01. (2019)
22. Tian, Z., et al.: Fcos: fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019)
23. Xu, Fengqiang, et al.: Scale-aware feature pyramid architecture for marine object detection. Neural Comput. Appl. **33**(8), 3637–3653 (2021)
24. Ghiasi, G., Tsung-Yi, L., Quoc, V.L.: Nas-fpn: learning scalable feature pyramid architecture for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
25. Lin, T.-Y., et al.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (2017)
26. Zheng, Z., et al.: Distance-IoU loss: faster and better learning for bounding box regression. In: Proceedings of the AAAI Conference on Artificial Intelligence vol. 34. No. 07 (2020)

27. Chen, Z., et al.: Piou loss: towards accurate oriented object detection in complex environments. In: European Conference on Computer Vision. Springer, Cham (2020)
28. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
29. Huang, G., et al.: Condensenet: an efficient densenet using learned group convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
30. Bochkovskiy, A., Chien-Yao, W., Hong-Yuan, M.L.: Yolov4: optimal speed and accuracy of object detection. arXiv:2004.10934 (2020)
31. Duan, K., et al.: Centernet: keypoint triplets for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019)
32. Rodner, E., Simon, M., Fisher, R., Denzler, J.: Fine-grained recognition in the noisy wild: sensitivity analysis of convolutional neural networks approaches. In: Procedings of the British Machine Vision Conference 2016. British Machine Vision Association (2016)

Xinbin Li received his M.Sc. degree in control theory and control engineering from Yanshan University, China, in 1999, and the Ph.D. degree in general and fundamental mechanics from Peking University, China, in 2004. He is now a professor in the Institute of Electrical Engineering, Yanshan University, China. His research interests include underwater acoustic networks and underwater image processing.

Haifeng Yu received his M.Sc. degree in control engineering from North China University of Science and Technology, China, in 2017, and the Ph.D. degree in control theory and control engineering of Yanshan University, China, in 2023. He is now a lecturer in the College of Electrical Engineering, North China University of Science and Technology, China. His research interests include image processing and deep learning.



Haiyang Chen received his B.S. degree in automation from Hebei University, China, in 2018. He is now a postgraduate in the Institute of Electrical Engineering, Yanshan University, China. His research interests include object detection.