Neural Computing and Applications (2022) 34:14881–14894
https://doi.org/10.1007/s00521-022-07264-8

ORIGINAL ARTICLE

Refined marine object detector with attention-based spatial pyramid pooling networks and bidirectional feature fusion strategy

Fengqiang Xu[2] · Huibing Wang[1] · Xudong Sun[1] · Xianping Fu[1,3]

Received: 15 August 2021 / Accepted: 29 March 2022 / Published online: 14 May 2022

Abstract
Marine object detection has become increasingly important in intelligent underwater robot. Because of color cast and blur in underwater images, features directly extracted from backbone networks usually lack interesting and discriminative characters, that affects performance on marine object detection. To this end, this paper proposes a novel refined marine object detector with attention-based spatial pyramid pooling networks and bidirectional feature fusion strategy to relieve the weakening of features and address marine object detection issues. Firstly, an attention-based spatial pyramid pooling network named as SA-SPPN is proposed to enrich interesting information and extend receptive field on original features extracted from backbone network. Based on enhanced multiple level features, the bidirectional feature fusion strategy is designed to fuse different level features and generate robust feature maps for detection. Specifically, the top-down connection could transfer semantic information from high-level features to enhance low-level features. The bottom-up pathway could extend resolution of high-level features. Furthermore, the cross-layer connections are integrated into both top-down passway and bottom-up passway to carry out multiple branch fusion. On bounding boxes regression phase, the distance-IoU loss is adopted to improve regression speed and accuracy. Finally, this paper conducts series experiments on underwater image datasets and URPC datasets to detect marine objects. The experimental results reveal that our approach could achieve impressive performance and reach 79.64% mAP on underwater image datasets, 79.31% mAP on URPC2019 datasets and 79.93% mAP on URPC2020 datasets, respectively. For standard object detection, the proposed algorithm also could realize notable performance and get 81.9% mAP on PASCAL VOC datasets.

Keywords Marine object detection · Feature enrichment · Feature fusion · Convolutional neural network

1 Introduction

With the development of underwater robot, marine object detection has become a hot and urgent research topic. Because it is the foundational condition for underwater robot to realize intelligent observation and automatic capture of marine objects. Detection algorithms based on underwater optimal image have superiority on real-time detecting small objects in short-distance detection task, taking holothurian and scallop as example.

However, marine object detection task based on underwater optimal image still faces great challenges on feature representation. Because of the scattering and absorption of light transferred under the water, underwater optimal images captured by underwater cameras are usually color cast and blurry, as shown in Fig. 1. Features directly extracted from underwater images with convolutional

✉ Xianping Fu
fxp@dlmu.edu.cn

Fengqiang Xu
fengqiang_xu@163.com

Huibing Wang
huibing.wang@dlmu.edu.cn

Xudong Sun
sxd@dlmu.edu.cn

[1] College of Information Science and Technology, Dalian Maritime University, Dalian 116026, China

[2] College of Software, Dalian Jiaotong University, Dalian 116028, China

[3] Peng Cheng Laboratory, Shenzhen 518055, China

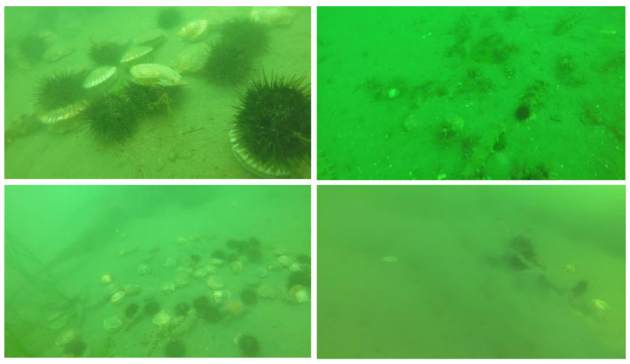


Fig. 1 Some frames in underwater datasets. Underwater optimal images captured by underwater robot are usually color cast and blurry

neural networks usually lack interesting and discriminative characters, that affects performance on marine object detection. Thus, some popular object detectors [1–12] are not effective when applied directly to marine object detection task. This paper summarizes this phenomenon as the weakening of features.

To deal with weakening of features, it is of great importance to reinforce original features extracted from backbone networks, as represented in Fig. 2. It mainly includes two way to reinforce features: feature enhancement and feature fusion. Recently, attention mechanism has been adopted in popular methods to enhance features, because it could focus on interesting features. There are lots of classic attention structure, such as [13–15], and so on. Woo et al. [15] applies attention-based feature refinement with two distinctive modules, channel and spatial, and improves representation power of CNN networks. So, this paper introduces spatial attention mechanism into our detector framework and develops an attention-based spatial pyramid pooling network to enrich features.

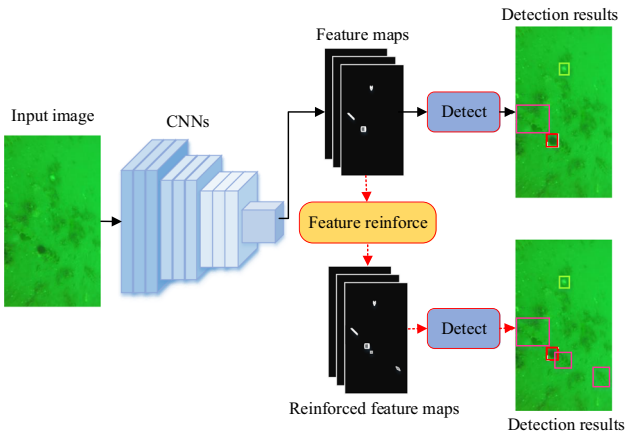


Fig. 2 The feature reinforcement strategy to relieve the weakening of features. By reinforcing feature maps extracted from CNNs, marine object detector could improve object detection performance

To further improve discrimination of features, a broad range of prior researches have been proposed in recent years. At the beginning, [1–5] just collect scale-fixed features generated by convolutional neural networks to detect object and cannot reach high accuracy. To adapt to different scale object detection, [6–9, 16, 17] extract different scale features from backbone networks. Recently, to further enrich features, [18] designs a top-down connection structure to carry out semantic information transferring from high-level features to low level. The main contribution of [18] is that it provides a novel strategy of feature fusion on different level features. Activated by [18, 19] proposes a bottom-up pathway to improve resolution of high-level features. Furthermore, [20] builds scalable feature pyramid network by neural architecture search. And [21] proposes an efficient feature fusion strategy. Based on above exploration, this paper designs a special bidirectional feature fusion architecture that could generate both high resolution and semantically strong features.

In this paper, we propose a novel refined marine object detector with attention-based spatial pyramid pooling networks and bidirectional feature fusion strategy. Firstly, to enhance original features extracted from backbone networks, we develop an attention based spatial pyramid pooling network to strengthen interesting information and extend receptive field of features. What's more, each feature generating branch adjoined to backbone network is integrated with SA-SPPN. Secondly, this paper designs a bidirectional feature fusion architecture to improve discrimination of features. On one hand, the top-down connection is adopted to enrich low-level features by fusing semantic information from high-level features. On the other hand, the bottom-up pathway is utilized to extend resolution of high-level features by fusing detail information from low-level features. Furthermore, this paper adds cross-layer fusion pathway into both vertical and horizontal path to provide multiple input features. Finally, this paper adopts distance-IoU loss to speed up bounding box regression. To validate performance of proposed method, we conduct experiments on underwater image datasets and reach 80.2% mAP. The experimental results reveal that our algorithm could improve performance on marine object detection.

The main contributions of this paper can be summarized as follows:

(1) An attention-based spatial pyramid pooling network is proposed to reinforce original convolutional features extracted from convolutional neural networks. SA-SPPN could increase the receptive field and separate out the most significant contextual features.

(2) A bidirectional feature fusion architecture is designed to strengthen the discriminative of feature maps. Our feature fusion manners include top-down up-sampling passway, bottom-up down-sampling passway and cross-layer fusion passway.
(3) The refined marine object detector is developed to improve performance on marine object detection. The experimental results reveal that our detector could achieve the latest state-of-the-art results on marine object detection.

The rest of the paper is organized as follows. Sect. 2 systematically introduces the proposed methods. Sect. 3 conducts experiments to support our method and analyzes experimental results. Sect. 4 summarizes related works involved in our algorithm. In addition, Sect. 5 makes a final conclusion on this paper.

2 The proposed method

To settle the issue on marine object detection, this paper proposes refined single shot detector with attention-based spatial pyramid pooling networks and bidirectional feature fusion strategy. In Sect. 2.1, we introduce the whole architecture of proposed method. In Sect. 2.2, we develop SA-SPPN structure to enhance features. In Sect. 2.3, we design bidirectional feature fusion network to build feature pyramid. In Sect. 2.4, we introduce distance-IoU loss for bounding box regression.

2.1 Framework architecture

Our object detector framework is mainly equipped with feature extraction, feature enrichment, feature fusion, and prediction head network. The architecture of proposed algorithm is represented in Fig. 3.

Firstly, we employ Darknet-53 as backbone network to extract original convolutional features from input images. Darknet is firstly proposed in [4], which has 24 convolutional layers followed by 2 fully connected layers. Then, [5] attempts various improvements on Darknet and proposes a new model, called Darknet-19, which has 19 convolutional layers and 5 maxpooling layers. Furthermore, [9] designs a new network named as Darknet-53, which is a hybrid approach between Darknet-19 and residual network stuff. Darknet-53 runs significantly faster than most detection methods with comparable performance. So, this paper adopts Darknet-53 as backbone network and extracts features from top three convolutional block to build feature pyramid.

Then, we develop an attention-based spatial pyramid pooling network equipped on each branch of backbone network to enhance interesting information and extend the receptive field of features. In SA-SPPN, we introduce spatial attention mechanism to adaptively refine intermediate feature map in spatial dimension. And spatial pyramid pooling network [22] could generate a fixed-length representation regardless of image size/scale and extend receptive field on convolutional features. This paper combines spatial attention mechanism with spatial pyramid pooling structure and redesigns them as a whole structure to enhance features. The details of SA-SPPN are introduced in Sect. 2.2.

Before predicting bounding boxes from features, we design an improved bi-directional feature pyramid network to fuse features from different layers and produce multi-scale refined features. After firstly proposed in [18], feature pyramid network becomes a crucial components in popular detection frameworks. Motivated by [18–21], this paper proposes a novel bidirectional feature fusion network to fuse features. The specific feature fusion manner is discussed in Sect. 2.3.

Based on final feature maps extracted from our architecture, prediction head could classify the bounding boxes to possible categories and regress them to the proper locations. In regression phase, we adopt the distance IoU loss function to speed up box regression process. The distance IoU loss has more specific regressing direction and could avoid unnecessary regression process.

2.2 Attention-based spatial pyramid pooling network for feature enrichment

This paper designs an attention-based spatial pyramid pooling network named as SA-SPPN, which is combined with spatial attention module and spatial pyramid pooling module. Recently, attention mechanism becomes popular in convolutional neural networks and shows good performance on computer vision tasks, such as classification, object detection, image translation, and so on. Attention not only tells where to focus, it also improves the representation of interests. Thus, this paper increases representation power of features by adopting attention mechanism. Spatial pyramid pooling component could increase the receptive field and generate the most significant context features without incurring extra computational burden. Original features extracted from convolutional layers are input into SA-SPPN and they will be reinforced. The output features are much discriminative.

To represent the whole process, input features of SA-SPPN are defined as $F_{\text{in}} \in \mathbb{R}^{c\times w\times h}$, and output features are defined as $F_{\text{out}} \in \mathbb{R}^{5c\times w\times h}$. Firstly, F_{in} is sent into spatial attention module to generate attention feature $F_a \in \mathbb{R}^{c\times w\times h}$. Spatial attention block is a pre-process

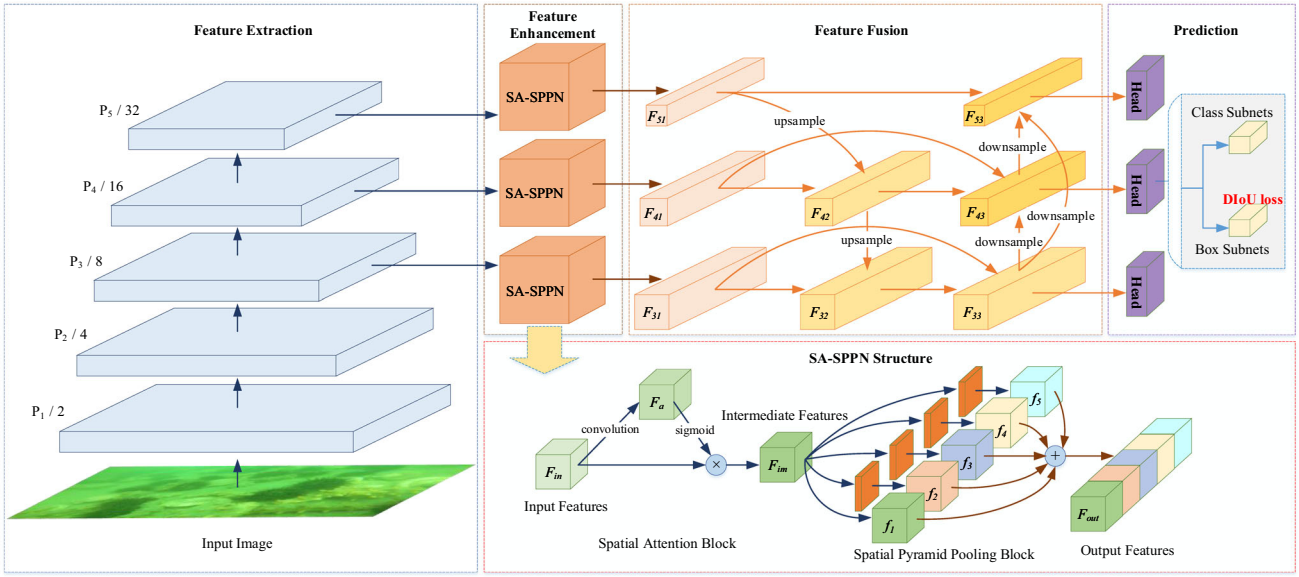


Fig. 3 The architecture of proposed marine object detector. We employ Darknet-53 as feature extraction network to get basic feature maps. Then, we design the attention-based spatial pyramid pooling network to enhance interesting features and augment receptive field of features. After that we build bidirectional feature fusion network to realize fast multi-scale feature fusion. Based on the refined feature maps, classification and regression are conducted to produce detection results. While regressing bounding boxes, we adopt the distance IoU loss to improve the speed of regression

component of SA-SPPN. Here, we replace max pooling and average pooling operation with convolutional operation to realize point-wise attention. So, F_a could be formulated as follows:

$$F_a = \mathrm{Conv}(F_{\mathrm{in}}), \tag{1}$$

where Conv is behalf of convolutional operation. F_a mainly includes interesting information extracted from F_{in}. Then point-wise addition operation between F_{in} and F_a is conducted to produce reinforced feature F_{im}, which is represented as follows:

$$F_{\mathrm{im}} = \sigma(F_a) \otimes F_{\mathrm{in}}, \tag{2}$$

where σ denotes sigmoid process and $\otimes$ represents the point-wise addition. Thus, F_{im} includes more interesting information than F_{in}, and will be input into SPP module as basic feature.

After then, SPP module augments receptive field of input feature map by series of max-pooling operation. For instance, one of augmented feature map could be formulated as follows:

$$f_2 = \mathrm{MaxPool}|_{s=1}^{5\times5}(F_{\mathrm{im}}), \tag{3}$$

where $\mathrm{MaxPool}|_{s=1}^{5\times5}$ represents max-pooling operation to generate $f_2 \in \mathbb{R}^{c\times w\times h}$. Here, the filter size is set as 5×5 and the mask strides by one pixel at each step. As shown in Fig. 4, we design a group of filter sizes $(1\times1, 5\times5, 9\times9, 13\times13, 17\times17)$ to conduct max-pooling operation on F_{im} and generate feature maps

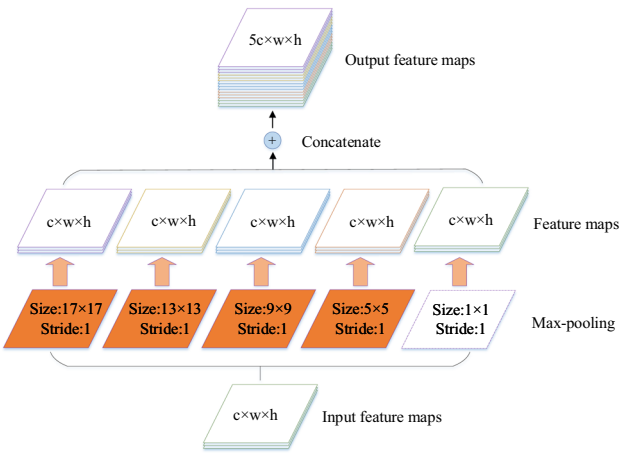


Fig. 4 Constructing of the spatial pyramid pooling network by concatenating feature maps from multiple branches. Each branch dilates receptive field of input feature maps with max-pooling operation. Specially, original feature maps could be regarded as operating with 1×1 max-pooling

$(f_1, f_2, f_3, f_4, f_5)$. Specifically, f_1 could be directly expressed by F_{im}. Thus, max-pooling operation with filter size of 1×1 could be omitted.

Finally, augmented feature maps are concatenated to output enhanced feature maps. The output feature maps could be represented as follows:

$$F_{\mathrm{out}} = f_1 \oplus f_2 \oplus f_3 \oplus f_4 \oplus f_5, \tag{4}$$

where $\oplus$ denotes the operation of feature concatenation. Each branch dilates receptive field with different scales.

After reinforced by SA-SPPN, feature maps extracted by backbone network are realized multi-scale receptive field augmentation. In this paper, we integrate SA-SPPN structure into each output branch of backbone network to enhance basic features.

2.3 Bidirectional feature fusion network to build feature pyramid

To enhance feature pyramidal representation, this paper proposes a novel bidirectional feature fusion network named as BiFFN with top-down fusion pathway and bottom-up fusion pathway. For multi-scale feature pyramidal representation, while high-level features are semantically strong but lower resolution, low-level features have richer detailed information but lack contextual content. Thus, recent research works are mainly focusing on generating feature representations that both high resolution and semantically strong. This section aims to optimize feature fusion strategy on feature pyramid network.

Activated by [18–21], this paper designs a special feature fusion architecture. As shown in Fig. 5, [18] combines two adjacent layers in feature hierarchy with top-down and lateral connections to enhance semantic information for low-level features. What's more, [19] adds an extra bottom-up pathway on feature pyramid to improve feature representations for lower resolution features. To improve model efficiency, [21] proposes several optimizations for cross-scale connections. Based on above researches, Our feature fusion architecture adopts both bottom-up pathway and top-down pathway to fuse features and adds cross-layer fusion pathway into both vertical and horizontal path to further fuse features.

As described in Fig. 5d, our feature fusion network includes three branch from P_3 to P_5. High-level features are up-sampled to enhance semantic information for low-level features with top-down pathway. after then, low-level features are down-sampled to improve resolutions and enrich detail information for high-level features by bottom-up pathway. Meanwhile, cross-scale connections could provide multiple input for feature fusion operation. So, our feature fusion network could fuse more features without adding much cost.

To represent the process of feature fusion, input feature maps from P_3 to P_5 are defined as F_{31}, F_{41} and F_{51}, respectively. The intermediate feature maps of F_{32} and F_{42} are formulated as follows:

$$\begin{cases} F_{42} = \mathrm{Conv}(F_{41} \oplus \mathrm{Resize}^{+}(F_{51})) \\ F_{32} = \mathrm{Conv}(F_{31} \oplus \mathrm{Resize}^{+}(F_{42})) \end{cases}, \tag{5}$$

where Resize^{+} denotes up-sampling function to increase the scale of features. Finally, the output feature maps are formulated as follows:

$$\begin{cases} F_{33} = \mathrm{Conv}(F_{31} \oplus F_{32}) \\ F_{43} = \mathrm{Conv}(F_{41} \oplus F_{42} \oplus \mathrm{Resize}^{-}(F_{33})) \\ F_{53} = \mathrm{Conv}(F_{51} \oplus \mathrm{Resize}^{-}(F_{32}) \oplus \mathrm{Resize}^{-}(F_{42})) \end{cases}, \tag{6}$$

where Resize^{-} denotes down-sampling function to reduce the scale of features.

After fusion process, output features are enhanced with semantic information and details from contextual layers. Therefore, the feature pyramid generated from BiFFN could perform well on prediction.

2.4 Distance-IoU loss for bounding box regression

Bounding box regression is crucial to object detection task. Although IoU loss [23] and generalized IoU loss [24] have been proposed to tailor to the IoU metric, they still suffer from the problems of slow convergence and inaccurate

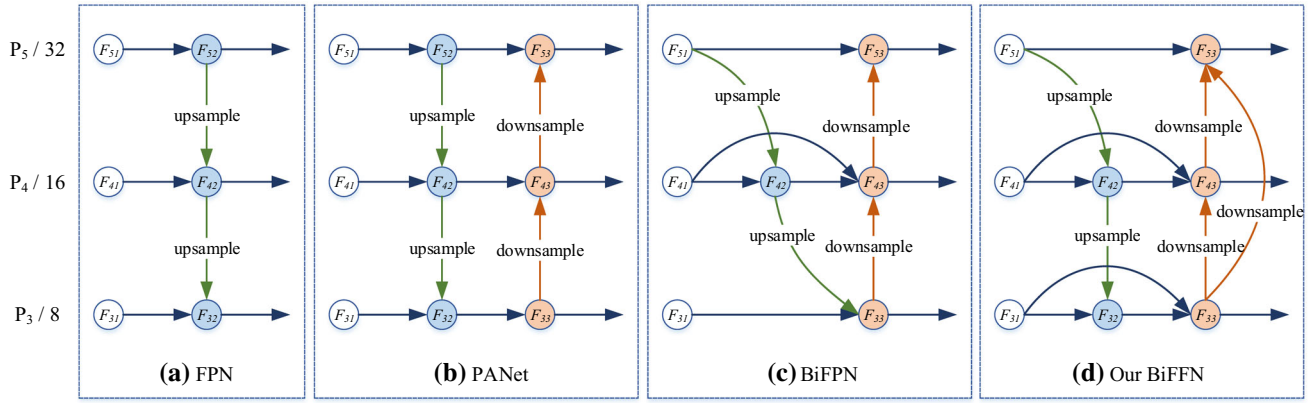


Fig. 5 The design of feature fusion network. **a** FPN [18] proposes a top-down pathway to fuse multi-scale features from low-level layers to high-level layers. **b** PANet [19] introduces a bottom-up pathway based on FPN. **c** BiFPN [21] adds cross-layer fusion pathway and omit some medial node. **d** Our BiFFN adds cross-layer fusion pathway into both vertical and horizontal path to fuse features

regression. This paper adopts Distance-IoU loss [25] by incorporating the normalized distance between predicted box and target box to accelerate bounding box regression in training.

The intersection over union (IoU) between predicted box and ground-truth box is calculated as following:

$$\mathrm{IoU} = \frac{|B \cap B^{\mathrm{gt}}|}{|B \cup B^{\mathrm{gt}}|}, \tag{7}$$

where $B^{\mathrm{gt}} = (x^{\mathrm{gt}}, y^{\mathrm{gt}}, w^{\mathrm{gt}}, h^{\mathrm{gt}})$ is the ground-truth box and $B = (x, y, w, h)$ is predicted boxs.

The DIoU loss is formulated as follows:

$$L_{\mathrm{DIoU}} = 1 - \mathrm{IoU} + \frac{\rho^2(b, b^{\mathrm{gt}})}{c^2}, \tag{8}$$

where b and b^{gt} denote the central points of B and B^{gt}, $\rho(\cdot)$ is the Euclidean distance, and c is the diagonal length of the smallest enclosing box covering the two boxes. DIoU loss could directly minimize the distance of two boxes to provide moving directions for bounding boxes, even when non-overlapping with target box. Thus, DIoU loss achieves faster convergence for predicted box and target box.

3 Experiments and analysis

In this section, we design several experiments on different image datasets to verify the performance of proposed method on object detection. We firstly conduct comprehensive experiments on our 4 category underwater image dataset. Then, we continue testing on 4 category URPC2019 and URPC2020, respectively. To further explore effectiveness of our method, we experiment on the 20 category PASCAL VOC datasets [26] and compare with popular detector. This paper adopts mean average precision (mAP) as evaluation criterion of accuracy. The experimental results represent the performance of our method on detection task.

3.1 Implementation details

This paper takes Darknet-53 as backbone networks and initializes the detector with parameters pre-trained on ImageNet1k classification set [27]. Generally, we train the detector with stochastic gradient descent (SGD) for 50 K iterations. The learning rate is initially set as 0.001, which is reduced by a factor of 10 at 40 K and 45 K iterations, respectively. In addition, the weight decay is set as 0.0005 and the momentum is set as 0.95 during training phase. All of the experimental results are implemented using a Nvidia GeForce GTX 1080 Ti GPU and cuDNN v7.6 and an Intel Core i7-6700K@4.00 GHz. To reduce computing burden, each image should be firstly resized to 608×608 and then input into our model.

3.2 Experiments on our underwater image datasets

Our underwater image datasets are built to explore the detection of marine objects. Specifically, it is mainly including 25,400 pictures with 4 categories: holothurian, echinus, scallop, and starfish. Part of images in our datasets are captured by our underwater robot in naturalistic ocean environment, and others are from videos on Internet. We have labeled them by ourselves. To validate the performance of proposed algorithm, we conduct series of experiments on underwater image datasets, including ablation study and comparison with other detectors.

3.2.1 Ablation study

In this section, we conduct several ablation experiments to verify the effect of each component in proposed algorithm. This paper takes Darknet-53 as backbone network and combines each component on it to improve performance. The experimental results are listed in Table 1.

In Table 1, the first row is the detection results of original method. Normal FPN structure is added into Darknet-53 and executed on underwater image datasets. This strategy could reach 76.11% mAP, which is set as baseline performance. Then, the proposed SA-SPPN components are combined into original method to enrich features. Experimental results from first two rows in Table 1 reveal that the proposed SA-SPPN could achieve 1.52% mAP improvement. What's more, this paper designs a bidirectional feature fusion network to replace original FPN to fuse contextual information. The comparison of second row with third row in Table 1 illustrates that our BiFFN could generate 1.23% mAP gains on detection. To explore the contribution of distance IoU loss, this paper conducts experiments on original method. Results from first row and fourth row represent that adopting distance IoU loss could get 0.7% mAP gains on marine object detection.

The last row in Table 1 is the setting of proposed algorithm in this paper, which combines with SA-SPPN components, designed BiFFN, and distance IoU loss. Experimental results show that our proposed method could reach 79.64% mAP on marine object detection task, which outperform original method by 3.53% mAP.

Some detection results of proposed method on underwater image datasets are represented in Figs. 6 and 7. Our method has good performance on marine object detection not only for big scale targets but also for small objects. Even in blurry environment, our algorithm still works well

Table 1 Ablation experiments on underwater image dataset

Darknet-53	SA-SPPN	FPN	BiFFN	DIoU	AP(%)				mAP(%)
					Holothurian	Echinus	Scallop	Starfish	
✓		✓			73.02	72.46	81.82	77.14	76.11
✓	✓	✓			74.71	74.22	82.91	78.68	77.63
✓	✓		✓		76.59	75.47	83.47	79.91	78.86
✓		✓		✓	73.74	73.05	82.60	77.85	76.81
✓	✓		✓	✓	**77**.22	**76**.38	**84**.23	**80.73**	**79**.64

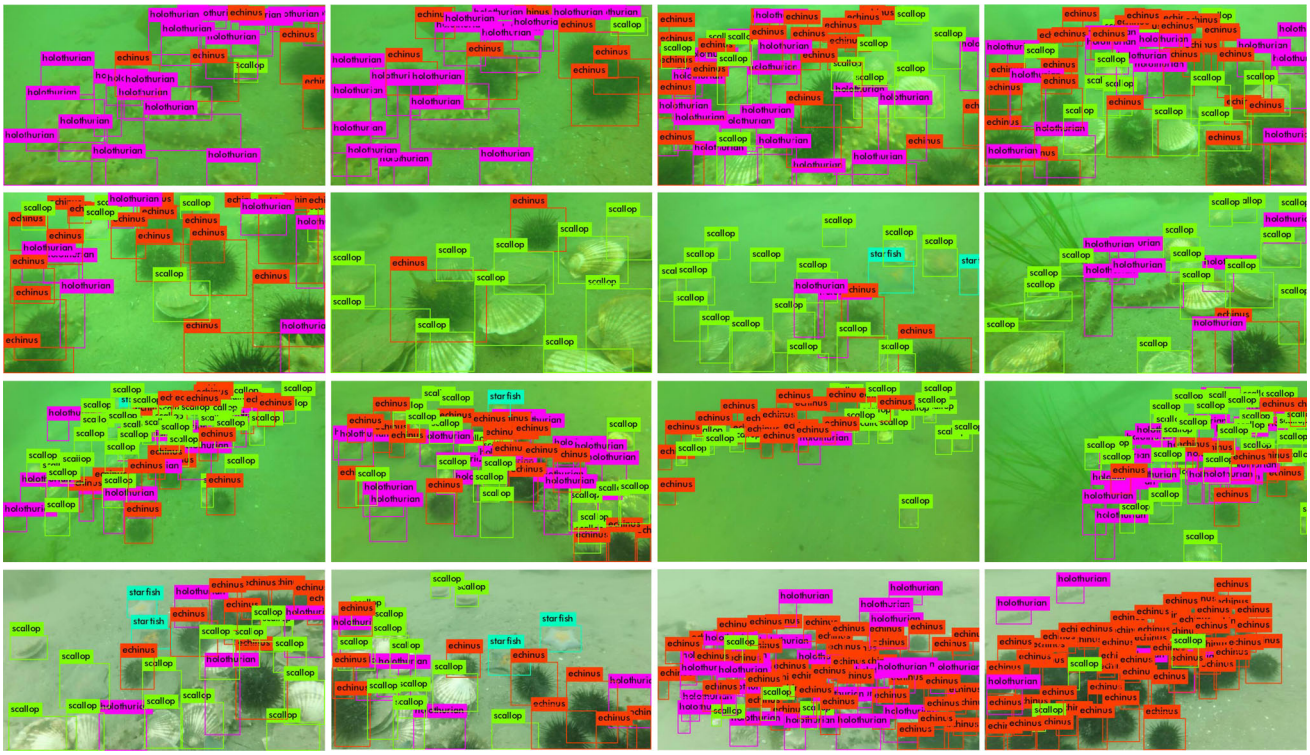


Fig. 6 Qualitative detection results of proposed algorithm on our underwater image dataset. Different categories of objects are drawn with different color

and could detect almost all targets. Nevertheless, our method still faces the challenge of missing detection, as some objects are difficult to be discriminated from background. For instance, third row in Fig. 7 is local area of second row images. There are some missing detected objects that are labeled with blue rectangle.

The training loss of our model is represented in Fig. 8. When the learning rate is reduced at 40K iterations during training phase, the loss decreases obviously. In addition, the Precision-Recall curves of different object categories on test image dataset are shown in Fig. 9. Different color curves represent different categories of objects.

3.2.2 Comparison with popular detector

To compare with popular detectors on marine object detection task, we conduct experiments with popular detectors using default settings in opened source code on underwater image datasets. And the experimental results are listed in Table 2.

Recent popular object detectors, such as Faster R-CNN, YOLO, SSD, and so on, have represented interesting performance on usual object detection task. However, it is still challengeable on marine object detection task. So this paper conducts experiments on underwater image datasets using popular detectors and collects results to compare. As shown in Table 2, while changing backbone network from ZFNet to VGGNet, Faster R-CNN could achieve 69.16%

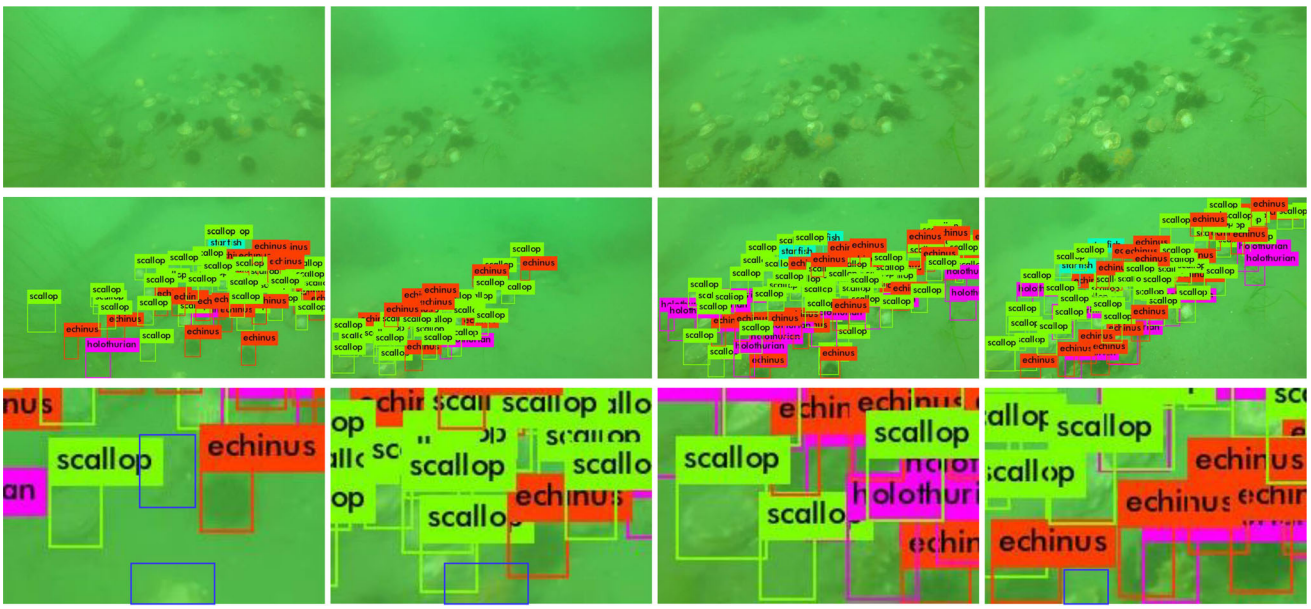


Fig. 7 Qualitative detection results of small objects in underwater images. The first row is behalf of original images and the second row represents detection results on original images. Some local areas in second row images are zoomed in and shown in third row. And the missing detected objects are labeled with blue rectangle

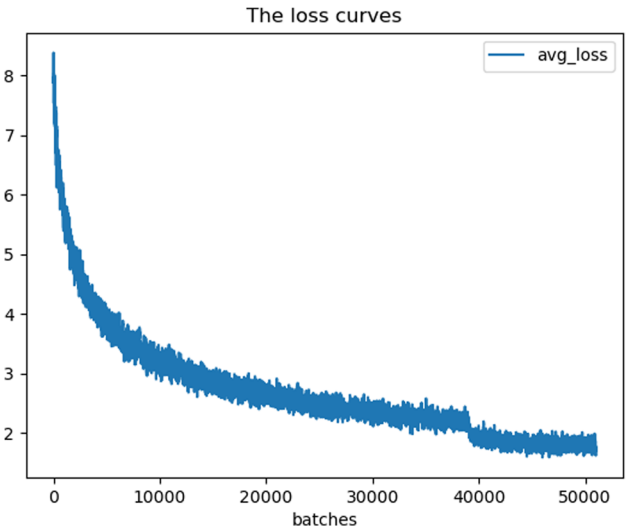


Fig. 8 The training loss of our model

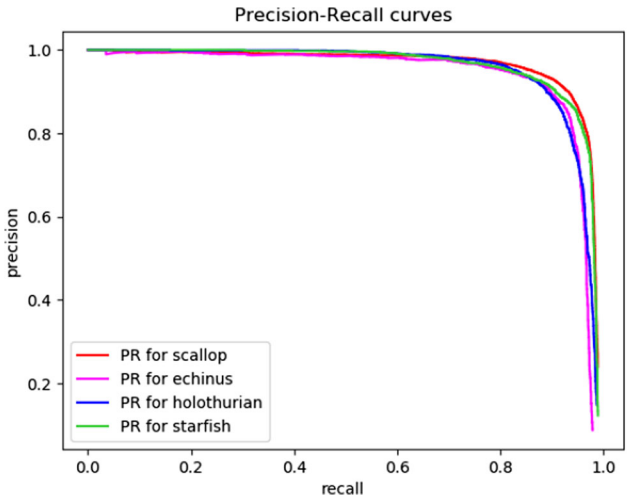


Fig. 9 The Precision-recall curves of different object categories on test image dataset

mAP. Comparing the results from third row to sixth row, it is surprising that YOLO series methods have continuous improvement on detection. At the beginning, first vision YOLO approach just could get 61.18% mAP with 41 FPS. But YOLOv4 has realized impressive performance of 79.26% mAP with 65 FPS. YOLOv5m could get competitive performance of 79.19% mAP with 68 FPS. The development of YOLO series methods is heuristic. In addition, SSD detector could obtain moderate precision with fast processing speed. Although FPN and SA-FPN methods acquire excellent performance on precision, they cost too much computing time. The experimental results reveal that our proposed method performs best on marine object detection task with 79.64% mAP and acceptable processing speed.

3.3 Experiments on URPC datasets

In this part, we evaluate our approach on two opened underwater datasets URPC2019 and URPC2020, which are from the Underwater Robot Picking Contest.[1] The URPC2019 and URPC2020 datasets have four object

[1] Underwater Robot Picking Contest. http://www.cnurpc.org/.

Table 2 Comparison with popular detectors on the underwater image datasets

Approach	Backbone	Input size	FPS	mAP(%)
Faster R-CNN	ZFNet	$\sim 1000 \times 600$	14	61.95
Faster R-CNN	VGGNet	$\sim 1000 \times 600$	5.6	69.16
YOLO	GoogLeNet	448×448	41	61.18
YOLOv2	Darknet-19	416×416	**61**	73.86
YOLOv3	Darknet-53	416×416	30	74.43
YOLOv4	CSPDarknet-53	512×512	65	79.26
YOLOv5m	CSPDarknet-53	640×640	68	79.19
SSD	VGGNet	300×300	42	70.03
FPN	ResNet-50	$\sim 1280 \times 768$	4.1	74.25
SA-FPN	ResNet-50	$\sim 1280 \times 768$	3.5	76.27
Ours	Darknet-53	608×608	29	**79**.64

categories, including echinus, scallop, holothurian and starfish.

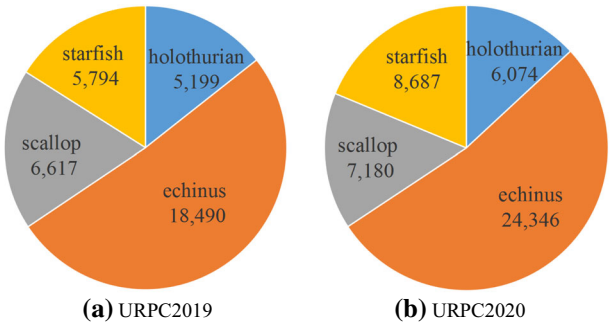
As represented in Table 3, the URPC2019 dataset has 4757 images, which are split into a training set of 3567 images and a testing set of 1190 images. The URPC2020 dataset has 6575 images, which are split into a training set of 4929 images and a testing set of 1646 images. What's more, we have finished statistics of ground-truth annotations of different categories on URPC2019 and URPC2020, respectively. Figure 10 represents that echinus is more ampler than other category objects and takes over half of annotations. Holothurian, scallop, and starfish have comparative ground-truth boxes.

This paper conducts experiments with proposed algorithm on URPC2019 and URPC2020 datasets, separately. The experimental results are listed in Table 4. Our algorithm could achieve 79.31% mAP on URPC2019 and 79.93% mAP on URPC2020. Notably, detection performance on echinus is higher than others, and holothurian is hard to detect in URPC datasets.

Some detection results of proposed method on URPC2019 dataset and URPC2020 dataset are shown in Figs. 11 and 12, respectively. Detection results reveal that the proposed method could perform well in different underwater conditions, even with complicated background.

Table 3 The training and testing images in URPC datasets

Datasets	Total images	Training images	Testing images
URPC2019	4757	3567	1190
URPC2020	6575	4929	1646



(a) URPC2019 **(b)** URPC2020

Fig. 10 The statistics of ground-truth boxes of different categories on URPC2019 and URPC2020. Each dataset has four categories

Table 4 Experimental results on URPC dataset

Datasets	AP(%)				mAP(%)
	Holothurian	Echinus	Scallop	Starfish	
URPC2019	69.60	87.70	79.03	80.92	79.31
URPC2020	71.03	88.03	79.38	81.27	79.93

For instance, detection results of last two rows in Fig. 12 show that our trained detector could successfully detect targets even in rocks.

In addition, the variable light within the images and the object distance also could affect detection results. While lacking enough light, the images are dark that increases the difficulty of distinguishing objects from background. From Figs. 11 and 12, it is revealed that the distance between objects and the distance between object and camera also could affect detection results. While the distance between objects is small, the objects are easy to be occluded by others, that may lead to miss detection. Furthermore, the smaller distance between object and camera is, the bigger scales of objects in images are. Usually, detections on small objects are more difficult than large objects.

3.4 Experiments on pascal VOC datasets

To further explore the effect of proposed algorithm on standard object detection task, this paper also implements experiments on the Pascal VOC dataset. Images in Pascal VOC dataset are annotated with 20 classes. We train the designed detector on the VOC 2007 and VOC 2012 trainval sets (16551 images), and test on the VOC 2007 test set (4952 images). The experimental results are represented in Table 5.

We compare our proposed algorithm with one-stage object detectors and two-stage detectors, respectively. Generally, object detection approaches are usually divided into one-stage detection methods and two-stage detection

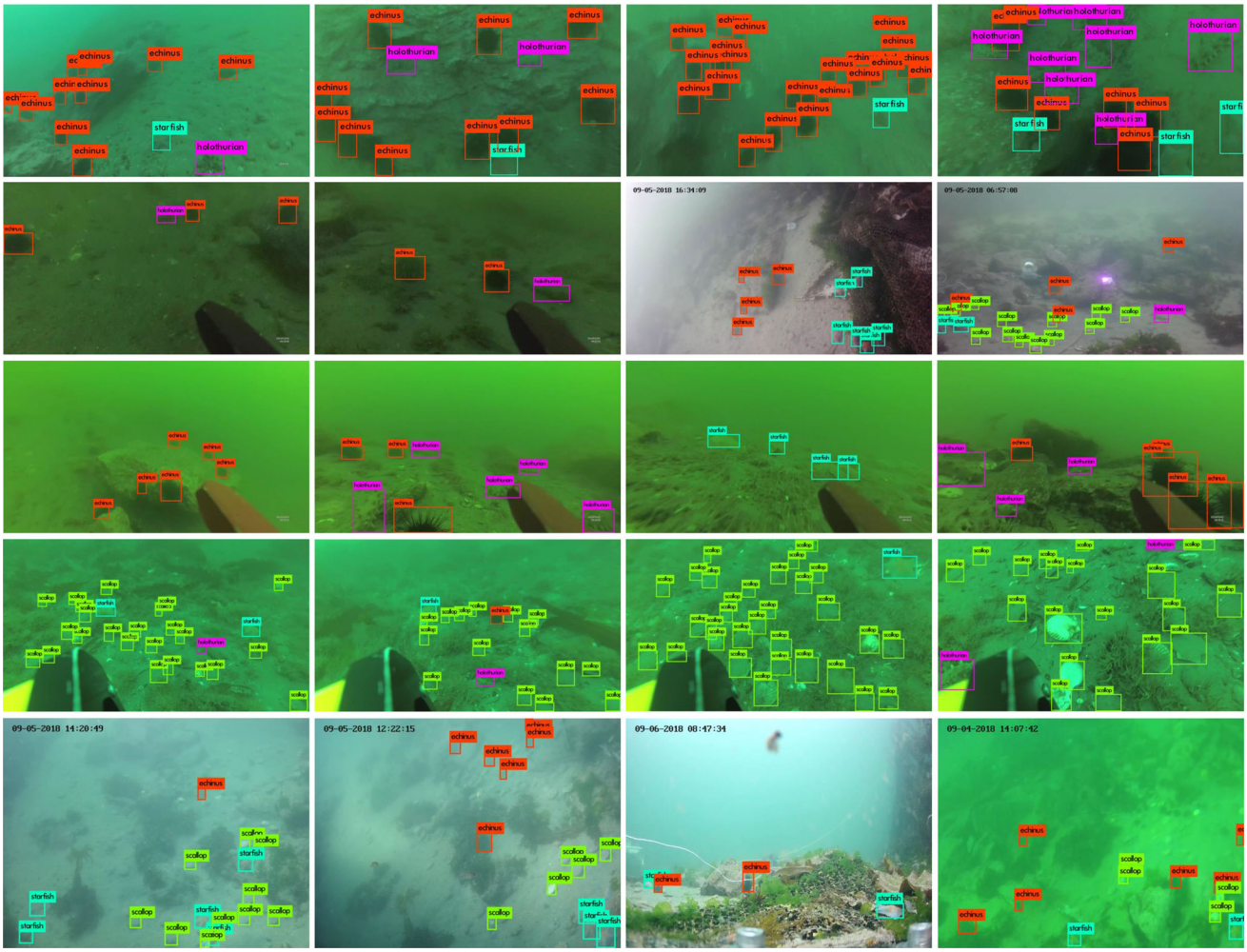


Fig. 11 Qualitative detection results of proposed algorithm on URPC2019 dataset

methods. While one-stage detectors could classify and detect targets with a single neural network, two-stage detectors need firstly generating region proposals with RPN and then detect objects based on proposals. Thus, one-stage detection approaches have advantages on real-time detection.

At the beginning, two-stage detectors could achieve surprising detection performance on precision. As shown in Table 5, faster R-CNN with VGGNet and ResNet-101 could reach 73.2% mAP and 76.4% mAP, respectively. FPN could get 77.1% mAP and SA-FPN can gain 79.1% mAP. However, the process of detection with two-stage detectors cost too much time. So it is challengeable for two-stage detectors to realize real-time detection. In contrast, one-stage methods could achieve fast detecting speed. In particular, YOLO series methods could process more than 34 frames per seconds. Recent YOLOv4 detector could achieve a competitive detection performance of 81.3% mAP with 65 FPS and YOLOv5m could reach 81.2% mAP with 68 FPS.. In addition, SSD, DSSD and DSOD methods also could realize reliable detection performance on the cost of increasing computation burden. Comparatively, the proposed algorithm could outperform state-of-the-art detectors and get 81.9% mAP on the PASCAL VOC datasets. The experimental results reveal that our designed framework also has good performance on standard object detection task.

4 Related work

4.1 Attention module

Attention plays an important role in human perception. Specifically, humans exploit a sequence of partial glimpses rather than a whole scene at once and selectively focus on salient parts in order to capture visual structure better [35]. For machine translation task, [36] proposes a sequence transduction model based entirely on attention, replacing the recurrent layers most commonly used in encoder-

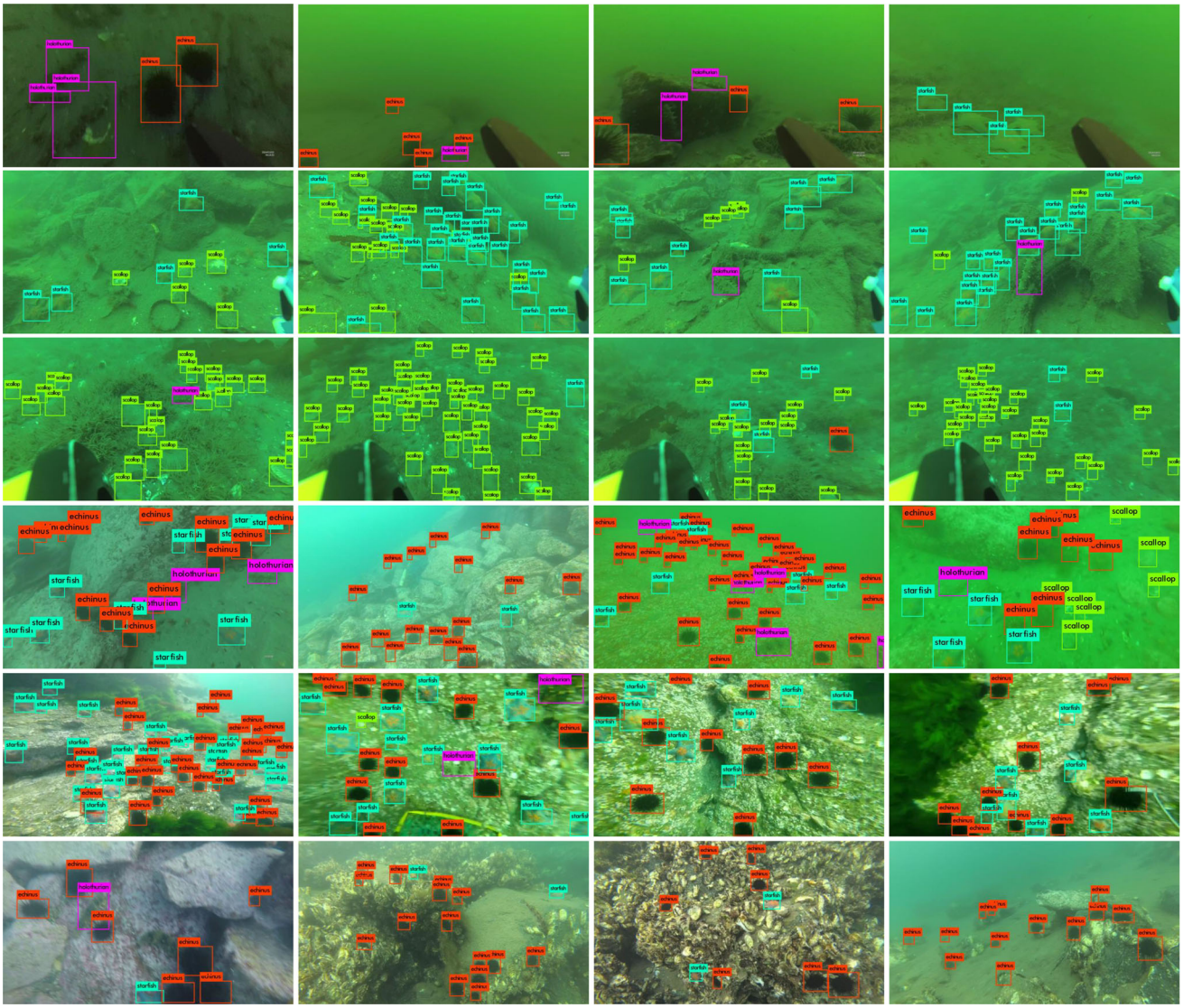


Fig. 12 Qualitative detection results of proposed algorithm on URPC2020 dataset

decoder architecture with multi-headed self-attention. Wang et al. [13] proposes a non-local blocks to capture long-range dependencies and bridges self-attention for machine translation to general task in computer vision, such as video classification, object detection and segmentation, pose estimation, and so on. To explore channel relationship, [14] proposes Squeeze-and-Excitation (SE) block to adaptively recalibrate channel-wise feature responses by explicitly modeling interdependencies between channels. [37] proposes residual attention network to generate attention-aware features. Woo et al. [15] applies attention-based feature refinement with two distinctive modules, channel and spatial, and improve representation power of CNN networks. Activated by [13] and [14, 38] simplifies non-local network and proposes the GC block to improve effectiveness. This paper adopts [15] as basic attention structure and modifies it from spatial-wise attention to point-wise attention. Specifically, [15] sequentially infers attention maps along two separate dimensions, channel and spatial, then the attention maps are multiplied to the input feature map for adaptive feature refinement. We replace max pooling and average pooling operations in spatial attention module with convolutional operation to realize point-wise attention.

4.2 Feature pyramidal representations

To detect multiple scale objects, it is of great importance to build and represent multi-scale features. In early works, [6, 7, 39] directly perform predictions based on the pyramidal feature hierarchy extracted from backbone networks. As one of the pioneering researches, [18] builds a feature

Table 5 Detection results on the PASCAL VOC 2007 datasets

Approach	Backbone	Input size	FPS	mAP(%)
Two-stage detectors				
Fast R-CNN [2]	VGGNet	$\sim 1000 \times 600$	0.6	70.0
Faster R-CNN [3]	VGGNet	$\sim 1000 \times 600$	7	73.2
Faster R-CNN [28]	ResNet-101	$\sim 1000 \times 600$	5	76.4
FPN [18]	ResNet-50	$\sim 1280 \times 768$	5	77.1
R-FCN [29]	ResNet-50	$\sim 1000 \times 600$	11	77.4
MR-CNN [30]	VGGNet	$\sim 1000 \times 600$	0.03	78.2
SA-FPN [31]	ResNet-50	$\sim 1280 \times 768$	4	79.1
One-stage detectors				
YOLO [4]	GoogLeNet	448×448	45	63.4
YOLOv2 [5]	Darknet-19	416×416	67	76.8
YOLOv3 [9]	Darknet-53	416×416	34	77.2
SSD300 [6]	VGGNet	300×300	46	74.3
SSD512 [6]	VGGNet	512×512	19	76.8
DSSD321 [7]	ResNet-101	321×321	9.5	78.6
DSOD300 [8]	DS/64-192-48-1	300×300	17.4	77.7
GFR-DSOD300 [32]	DS/64-192-48-1	300×300	17.5	78.9
YOLOv4 [33]	CSPDarknet-53	512×512	65	81.3
YOLOv5m [34]	CSPDarknet-53	640×640	68	81.2
Ours	Darknet-53	608×608	30	**81**.9

pyramid network (FPN) with a top-down pathway to transmit contextual information. Based on FPN, [19] proposes an extra bottom-up path aggregation network to enhance the entire feature hierarchy with accurate localization signals in lower layers. Ghiasi et al. [20] adopts neural architecture search and discovers a new feature pyramid architecture named as NAS-FPN, which consists of a combination of top-down and bottom-up connections to fuse features across scales. Although NAS-FPN achieves better accuracy, it requires thousands of GPU hours during search. To optimize multi-scale feature fusion with more intuitive and principled way, [21] proposes a weighted bidirectional feature pyramid network (BiFPN), which allows easy and fast multi-scale feature fusion. Wang et al. [40] directly handles the multi-view feature representation in the kernel space, which provides a feasible channel for direct manipulations on multiview data with different dimensions. Based on above researches, this paper aims to further explore the possibility of multi-scale feature fusion and designs a novel bidirectional feature fusion architecture.

5 Conclusion

This paper proposes a novel refined marine object detection framework with attention-based spatial pyramid pooling networks and bidirectional feature fusion strategy to address marine object detection issue. To verify the effectiveness of proposed approach, we conduct series experiments on underwater image datasets and URPC datasets. With the foundation of original features extracted from backbone network, an attention-based spatial pyramid pooling network named as SA-SPPN is designed to enrich interesting information and extend receptive field on original features. The experimental results reveal that introducing SA-SPPN could gain about 1.52% mAP improvement on marine object detection. Furthermore, this paper proposes bidirectional feature fusion strategy to fuse different level features from SA-SPPN branches. The output feature maps are discriminative and expressive. By ablation experiments, our new feature fusion strategy could improve 1.23% mAP. In addition, this paper adopts Distance-IoU loss to improve speed and accuracy of regression that could bring 0.7% mAP increase. Finally, our proposed algorithm achieves 79.64% mAP on underwater image datasets, 79.31% mAP on URPC2019 datasets and 79.93% mAP on URPC2020 datasets, respectively. Even on PASCAL VOC datasets, the designed approach could outperform state-of-the-art detectors and reach 81.9% mAP.

Our research work could achieve competitive performance on marine object detection task but still has room for further improvement. In the future, we plan to explore how to improve the speed of detection and integrate our refined marine object detector into underwater robot to realize fast and accurate detection.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China Grant 62176037 and 62002041, by Liaoning Revitalization Talents Program XLYC1908007, by the Dalian Science and Technology Innovation Fund 2021JJ12GX028, by the Liaoning Doctoral Research Start-up Fund Project Grant 2021-BS-075.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 580–587
2. Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 1440–1448
3. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. In: *Advances in neural information processing systems*, pp 91–99
4. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788
5. Redmon J, Farhadi A (2017) Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7263–7271
6. Wei L, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Alexander CB (2016) Ssd: single shot multibox detector. European conference on computer vision. Springer, New York, pp 21–37
7. Fu C-Y, Liu W, Ranga A, Tyagi A, Berg Lexander AC (2017) Dssd: deconvolutional single shot detector. arXiv preprint arXiv:1701.06659
8. Shen Z, Liu Z, Li J, Jiang Y-G, Chen Y, Xue X (2017) Dsod: learning deeply supervised object detectors from scratch. In: Proceedings of the IEEE international conference on computer vision, pp 1919–1927
9. Redmon J, Farhadi A (2018) Yolov3: an incremental improvement. arXiv preprint arXiv:1804.02767
10. Ma X, Jia W, Xue S, Yang J, Zhou C, Sheng QZ, et al (2021) A comprehensive survey on graph anomaly detection with deep learning. IEEE Trans Knowl Data Eng
11. Liu F, Xue S, Wu J, Zhou C, Hu W, Paris C, Nepal S, Yang J, Yu PS (2020) Deep learning for community detection: progress, challenges and opportunities. arXiv preprint arXiv:2005.08225
12. Su X, Xue S, Liu F, Wu J, Yang J, Zhou C, Hu W, Paris C, Nepal S, Jin D, et al (2021) A comprehensive survey on community detection with deep learning. arXiv preprint arXiv:2105.12584
13. Wang X, Girshick R, Gupta A, He K (2018) Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7794–7803
14. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)
15. Woo S, Park J, Lee J, Kweon SI (2018) Cbam: convolutional block attention module. pp 3–19
16. Wang H, Peng J, Zhao Y, Fu X (2020) Multi-path deep cnns for fine-grained car recognition. IEEE Trans Vehic Technol 99:1
17. Wang H, Peng J, Chen D, Jiang G, Zhao T, Fu X (2020) Attribute-guided feature learning network for vehicle re-identification. IEEE MultiMedia 27(4):112–121
18. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2117–2125
19. Liu S, Qi L, Qin H, Shi J, Jia J (2018) Path aggregation network for instance segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8759–8768
20. Ghiasi G, Lin T-Y, Le QV (2019) Nas-fpn: learning scalable feature pyramid architecture for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7036–7045
21. Tan M, Pang R, Le QV (2020) Efficientdet: scalable and efficient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10781–10790
22. He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 37(9):1904–1916
23. Yu J, Jiang Y, Wang Z, Cao Z, Huang T (2016) Unitbox: an advanced object detection network. In: Proceedings of the 24th ACM international conference on multimedia, pp 516–520
24. Rezatofighi H, Tsoi N, Gwak JY, Sadeghian A, Reid I, Savarese S (2019) 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 (CVPR)
25. Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D (2020) Distance-iou loss: faster and better learning for bounding box regression. In: The thirty-fourth AAAI conference on artificial intelligence, AAAI 2020, the thirty-second innovative applications of artificial intelligence conference, IAAI 2020, the tenth AAAI symposium on educational advances in artificial intelligence, EAAI 2020, New York, NY, USA, February 7–12, 2020, pp 12993–13000. AAAI Press
26. Everingham M, Van Gool L, Williams Christopher KI, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. Int J Comput Vis 88(2):303–338
27. Russakovsky O, Deng J, Hao S, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M et al (2015) Imagenet large scale visual recognition challenge. Int J Comput Vision 115(3):211–252
28. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
29. Dai J, Li Y, He K, Sun J (2016) R-fcn: object detection via region-based fully convolutional networks. In: Advances in neural information processing systems, pp 379–387
30. Gidaris S, Komodakis N (2015) Object detection via a multi-region and semantic segmentation-aware cnn model. In: Proceedings of the IEEE international conference on computer vision, pp 1134–1142
31. Fengqiang X, Wang H, Peng J, Xianping F (2021) Scale-aware feature pyramid architecture for marine object detection. Neural Comput Appl 33(8):3637–3653
32. Shen Z, Shi H, Yu J, Phan H, Feris R, Cao L, Liu D, Wang X, Huang T, Savvides M (2017) Improving object detection from scratch via gated feature reuse. arXiv: 1712.00886
33. Bochkovskiy A, Wang CY, Liao H (2020) Yolov4: optimal speed and accuracy of object detection
34. Jocher G, et al (2021) yolov5. *https://github.com/ultralytics/yolov5*
35. Larochelle H, Hinton G (2010) Learning to combine foveal glimpses with a third-order boltzmann machine. In: Proceedings of the 23rd international conference on neural information

processing systems, volume 1, NIPS'10, pp 1243–1251, Red Hook, NY, USA, Curran Associates Inc
36. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN , Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
37. Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, Wang X, Tang X (2017) Residual attention network for image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3156–3164
38. Cao Y, Xu J, Lin S, Wei F, Hu H (2019) Gcnet: non-local networks meet squeeze-excitation networks and beyond. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV) workshops
39. Cai Z, Fan Q, Feris RS, Vasconcelos N (2016) A unified multi-scale deep convolutional neural network for fast object detection. In: Bastian L, Jiri M, Nicu S, Max W (eds) Computer vision: ECCV 2016. Springer, Cham, pp 354–370
40. Wang H, Wang Y, Zhang Z, Fu X, Wang M (2019)Kernelized multiview subspace analysis by self-weighted learning