Neural Computing and Applications (2021) 33:3637–3653
https://doi.org/10.1007/s00521-020-05217-7

ORIGINAL ARTICLE

Scale-aware feature pyramid architecture for marine object detection

Fengqiang Xu[1] · Huibing Wang[1] · Jinjia Peng[1] · Xianping Fu[1,2]

Received: 26 December 2019 / Accepted: 17 July 2020 / Published online: 30 July 2020

Abstract

Marine object detection is an appealing but challengeable task in computer vision. Even though recent popular object detection algorithms perform well on common classes, they cannot acquire satisfied detection performance on marine objects because underwater images are affected by color cast and blur, and scales of the target in underwater images are usually small. These phenomena aggravate the difficulty of detection. Thus, it is urgent to design a proper structure to settle marine object detection issues. To this end, this paper proposes a novel scale-aware feature pyramid architecture named SA-FPN to extract abundant robust features on underwater images and improve the performance on marine object detection. Specifically, we design a special backbone subnetwork to improve the ability of feature extraction, which could provide richer fine-grained features for small object detection. What is more, this paper proposes a multi-scale feature pyramid to enrich the semantic features for prediction. Each feature map is enhanced by the higher level layer with context information through a top-down upsampling pathway. Through obtaining ample feature maps on underwater images, our algorithm could generate multiple bounding boxes for each target. To mitigate the reduplicative boxes and avoid miss suppression, we replace the non-maximum suppression method with soft non-maximum suppression. In this paper, we evaluate our algorithm on underwater image datasets and achieve 76.27% mAP. Meanwhile, we conduct experiments on PASCAL VOC datasets and smart unmanned vending machines datasets and get 79.13% mAP and 91.81% mAP, respectively. The experimental results reveal that our approach achieves best performance not only on marine object detection, but also on common classes.

Keywords Marine object detection · Feature pyramid network · Non-maximum suppression · Underwater image

1 Introduction

Marine object detection is a tricky but crucial task in computer vision. It is the foundation of ocean exploration and marine object intelligent detection. Because of the urgent demand in underwater robot developments, marine object detection task has drawn an appealing attention in recent years. It is the precondition for underwater robot to realize automatic capture. Although object detection has achieved success in common class datasets, marine object detection task still faces great challenges.

In recent years, the popular object detection approaches [3–7], based on convolutional neural networks (CNNs), have obtained good performance on common classes. However, these methods are not effective when applied directly to marine object detection task because underwater images captured by underwater cameras have poor visibility as shown in Fig. 1; this results from the scattering and absorption of light transferred under the water [8–11]. Specifically, underwater images are much blur than the ones captured out of water within the same distance and are deeper green, called color cast [12]. That leads to disappearance of fine-grained information on marine target in underwater images. What is worse, marine objects have protective coloration and aggregation effect. Thus, targets

✉ Xianping Fu
fxp@dlmu.edu.cn

Fengqiang Xu
xfq@dlmu.edu.cn

Huibing Wang
huibing.wang@dlmu.edu.cn

Jinjia Peng
jinjiapeng@dlmu.edu.cn

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

[2] Peng Cheng Laboratory, Shenzhen 518055, China

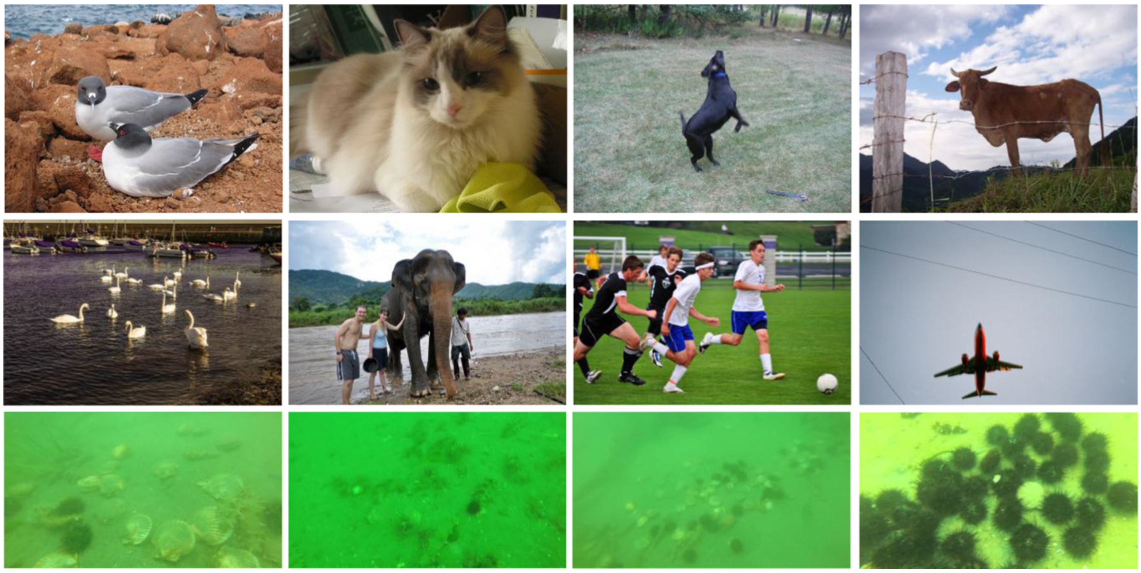


Fig. 1 The comparison of datasets. The first row and second row are images from the PASCAL VOC datasets [1] and Microsoft COCO datasets [2], respectively. The last row shows images from the underwater image datasets. Comparatively speaking, underwater images are blur and color cast. And the scales of marine objects are small. What is worse, marine objects have protective coloration and tend to live together

in underwater images are usually crowded and have small scales that aggravate the challenge of marine object detection task. So, it is necessary to explore and propose a special framework to solve marine object detection issue.

For CNNs, different level of convolutional layer extracts different scale features [13]. While the lower level layer could extract abundant fine-grained features, the higher level layer mainly focuses on semantic features [14]. For larger object, which is divided according to the relative size to whole image, the semantic features have effective contributions on detection task. However, the fine-grained characteristics provide crucial distinction for small object detection [15]. For marine object detection, it is important to build a multi-scale features, which include not only abundant fine-grained features but also strong semantic features.

Popular detectors in [3–6] just take the final feature maps to detect target. Liu et al. [7] uses multi-scale convolutional bounding box outputs attached to multiple feature maps at the top of the network. Fu et al. [16] introduces additional large-scale context with a deconvolution module. Recently, feature pyramid network (FPN) [17] is exquisite model architecture to generate pyramidal feature representations for object detection, which is popularly adopted by current object detection frameworks [18–23]. It adopts ResNets [24] to extract different scale features and design the bottom-up pathway, the top-down pathway and lateral connections to fuse the features. The top-down pathway upsamples spatially coarser, but semantically stronger, feature maps from higher pyramid levels. Meanwhile, these features are enhanced with features from the bottom-up pathway via lateral connections. Feature pyramid network provides potential feature maps that could be adopted as the fundamental feature structure to build the special feature architecture.

In this paper, we propose a novel scale-aware feature pyramid architecture based on FPN to detect marine objects. Firstly, we propose a special backbone subnetwork combined with a stacked convolutional layers. Each layer convolutes on input images with a small-scale filters and reserves abundant fine-grained information. This information is crucial to feature maps extraction. Secondly, we build a multi-scale feature pyramids. Different feature maps in our pyramids are generated by different convolutional blocks. What is more, the lower level feature maps, that have accurate location but weak semantics, are enhanced with strong semantic features from higher level by upsampling pathway. The enhanced feature pyramids are adopted to predict targets. To suppress reduplicative bounding boxes of each object, this paper takes soft non-maximum suppression (Soft-NMS) method to eliminate duplicates and solve miss suppression issue result by non-maximum suppression (NMS). Above all, the proposed algorithm improves performance on marine object detection task, especially on marine object detection.

The major contributions of this paper are summarized as:

(1) We propose a novel scale-aware feature pyramid architecture to execute marine object detection task.

Our structure improves the ability on feature extraction and performs well on marine object detection.

(2) We propose a backbone subnetwork structure to extract abundant fine-grained features. The first convolution layer of original ResNet-50 is replaced with a three-stacked convolution block. Fine-grained features are discriminative that benefit for small object detection.

(3) We propose a novel multi-scale feature pyramid to enrich semantic feature maps. Our feature pyramid is combined with several different scale feature maps. Each feature map is enhanced by the higher level through a top-down upsampling path. This structure could reinforce the features with context information and strengthen the discrimination of feature maps.

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

2 Related work

2.1 Object detection

Object detection is a heavily researched topic in computer vision, such as vehicle detection, pedestrian recognition, and autonomous driving. There has been a large body of researches on object detection with deep learning. According to whether region proposal is needed or not, popular object detection methods based on CNN mainly include region proposal-based methods and proposal-free methods.

Proposal-based methods [3–5, 25, 26] achieve excellent object detection accuracy. They mainly cover two stages: (1) they firstly generate region proposal based on feature maps, and (2) then they classify the proposal as specific category and produce accurate location for each object. Computational cost is the bottleneck of these approaches. Furthermore, Dai et al. [27] propose position-sensitive score maps to address a dilemma between translation-invariance in image classification and translation-variance in object detection. He et al. [28] extend faster R-CNN by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition.

Proposal-free paradigms [6, 7, 16, 29–31] principally focus on realizing real-time detection. These methods frame object detection as a regression problem using a single neural network to detect object and category from full images in one evaluation. So, it can be optimized end-to-end directly on detection performance. In addition, Shen et al. [32, 33] explore training object detectors from scratch without pretraining and contribute a set of design principles.

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. Girshick et al. [3], Girshick [4], Ren et al. [5] and Redmon et al. [6] just take the final feature maps to detect target. Long et al. [34] and Hariharan et al. [35] sum partial scores for each category over multiple scales to compute semantic segmentations. Liu et al. [36] and Kong et al. [37] concatenate features of multiple layers before computing predictions. Liu et al. [7] adds convolutional feature layers to the end of the truncated base network and produces multi-scale feature maps for detection. Fu et al. [16] propose a deconvolution module to introduce additional large-scale context in object detection. Lin et al. [17] designs a pyramid architecture to extract multiple feature maps from different layers. Lin et al. [19] and Tian et al. [20] adjust the feature maps on FPN and take higher level feature maps to predict object. In this paper, We take FPN as a baseline and build our multi-scale feature pyramid.

2.3 Non-maximum suppression

Non-maximum suppression is a necessary component employed in state-of-the-art object detection method. As it could distinguish the detections as positive or negative examples by computing overlap between each pair of detection boxes and merge all detections that belong to the same object. The method widely adopted in object detections [3–7, 16, 32] is described as greedy NMS, as it selects a bounding box with the maximum detection score for the object and suppress its neighboring boxes using a predefined overlap threshold.

Greedy NMS method has shortcoming on miss elimination, so series of improved approaches are proposed recently. Rothe et al. [38] present a clustering-based NMS algorithm based on affinity propagation. Hosang et al. [39] propose a convent designed to perform NMS of a given set of detections, which could overcome the intrinsic limitations of greedy NMS and obtain better recall and precision. Hosang et al. [40] propose a new network architecture designed to perform NMS, using only boxes and their score. Bodla et al. [41] propose Soft-NMS that decays the detection scores of all other objects as a continuous function of their overlap. Thus, the eliminated boxes in greedy NMS have the chance to be selected for other objects. Based on the discovery that the probabilities for class labels naturally reflect classification confidence, and localization confidence is absent, Jiang et al. [42] propose IoU-guided NMS procedure to take the localization confidence into account. He et al. [43] propose a novel bounding box regression loss for learning bounding box transformation and localization variance together, which helps to merge neighboring bounding boxes during NMS.

3 Scale-aware feature pyramid architecture

3.1 Model architecture

To settle the issue on marine object detection, this paper proposes scale-aware feature pyramid algorithm, and the model architecture is represented in Fig. 2. Our whole method could be concluded as three process: feature extraction, region proposal, and object detection.

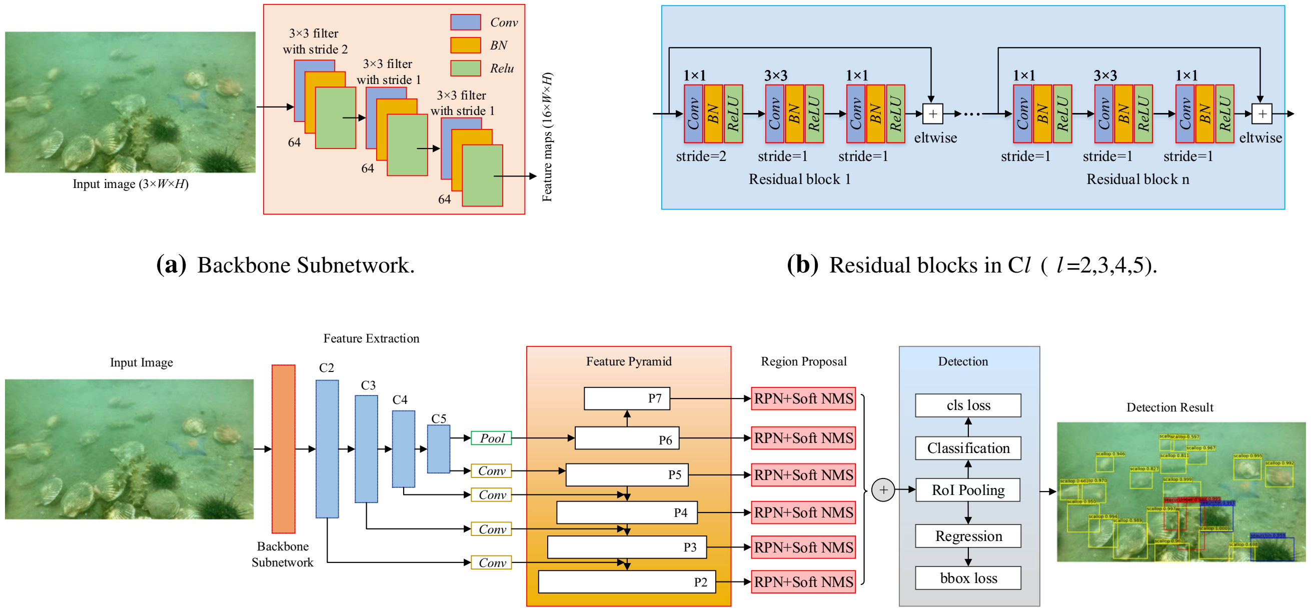
3.1.1 Feature extraction

Feature extraction is the foundational but crucial process in marine object detection. Because of the scattering and absorption of light transferred under the water, underwater images usually have poor visibility, which results in disappearance of details feature. This paper proposes feature extraction architecture based on residual network to obtain abundant and robust feature maps.

Firstly, we propose backbone subnetwork by replacing the first convolutional layer with three-stacked convolutional blocks. For the original structure in ResNet-50, first convolutional layer processes input images with a 7×7 size of filter that may weaken some fine-grained information. This paper takes three-stacked convolutional blocks to extract abundant fine-grained features. Each layer in block contains a smaller size of filter. So the generated features are discriminative for small object detection.

Then, several residual blocks are following with backbone subnetwork as feature extraction network. Each block designs with residual thoughts that provide two pathway to transfer parameters. One pathway processes with several convolutional layers and the other escapes from it. There are different scale feature maps generated by these residual blocks.

Finally, this paper builds a multi-scale feature pyramid based on feature extraction network. Each low level feature map is enhanced with context information from top-down pathway. Specifically, half of low level features are acquired from low level residual blocks and the other half are upsampled from higher level feature map. What is more, our pyramidal feature maps contain higher level



(a) Backbone Subnetwork.

(b) Residual blocks in Cl (l=2,3,4,5).

(c) The whole architecture of proposed scale-aware feature pyramid network.

Fig. 2 The architecture of scale-aware feature pyramid networks. **a** The structure of backbone subnetwork. **b** The basic residual blocks adopted in convolutional block C l ($l = 2, 3, 4, 5$). **c** The structure of proposed scale-aware feature pyramid network. We adopt ResNet as feature extraction network and design a particular backbone structure to generate abundant fine-grained features. To enrich semantic information, we design a multi-scale feature pyramid structure. After extracting ample feature maps, each target could generate several bounding boxes. To suppress reduplicative boxes, it is essential to replace traditional NMS algorithm with Soft-NMS. Finally, R-CNN is concatenated to calculate the loss of classification and regression

features. This pyramidal feature maps contain abundant fine-grained and strong semantic information that benefit for marine object detection.

3.1.2 Region proposal

After acquiring feature maps, this paper generates proposal bounding box by region proposal networks(RPN). We adapt RPN by replacing the single-scale feature map with our multi-scale feature maps. RPN produces multiple scale bounding boxes for each pixel on feature maps and suppresses the reduplicative boxes. However, classical non-maximum suppression method has issue of miss elimination. Thus, this paper replaces NMS method with Soft-NMS algorithm to release miss suppression issue. As a result, the proper proposals could be selected by RPN.

3.1.3 Object detection

While the proposals are produced, this paper detects targets with fast r-CNN algorithm. ROI data map with feature maps by ROI pooling layer and different scale ROI match with different scale feature map. Furthermore, the reduplicative boxes are suppressed by soft-NMS method, and each box is detected as a specific category with locations.

3.2 Backbone subnetwork for abundant fine-grained features extraction

Fine-grained features are essential to small object detection, which contain discriminative information, such as fine-grained texture and edge information. Abundant fine-grained features are beneficial to distinguish the target from the similar but inhomogeneous one. For instance, the scallop could be distinguished from some stone with abundant fine-grained texture features. However, the sea cucumber may be ignored without fine-grained edge features. So extracting abundant fine-grained features is the foundation of object detection task.

For CNNs in object detection, first convolution layer is the foundation of following net structures because it is responsible for extracting detail features from input images. However, due to the constraint of computation speed and memory capacity, filters in first convolution layer are usually designed as lager scale. This design may result in disappearance of fine-grained features because it is easily affected by the surrounding noises and has the disadvantage of subtle feature reservation, especially for small objects. To this end, we analyze the structure of ResNet and VGGNet and propose backbone subnetwork called Root-ResNet to extract abundant fine-grained features.

Our backbone subnetwork is designed based on the ResNet-50. As described in Fig. 3b, the first convolution layer of original ResNet-50 is defined as 7×7 kernel size with stride 2. This setting is easily affected by the surrounding noises and may lead to disappearance of detail features for small objects. Inspired by [32, 44], we propose a backbone subnetwork, which replaces the first convolution layer of ResNet-50 with stacked convolutional block, to improve the competence of feature extraction. As shown in Fig. 3c, our backbone subnetwork constitutes with three-stacked 3×3 convolution block, where the stride size of the first convolution layer is set as 2 and the other layers as 1. Each convolution layer is adjacent to the BatchNorm layer and ReLU layer, which could optimize the parameters. By convoluting on input images with stacked small-scale filters, our method could alleviate the affection of surrounding noises and acquire abundant fine-grained information. In Fig. 3d, we further change the stride size of first 3×3 convolution layer from 2 to 1. Without downsampling operation, the detection performance has been improved slightly because it is able to exploit more detail information from the images, so as to extract powerful fine-grained features for small object detection.

3.3 Multi-scale feature pyramid for semantic information enrichment

To acquire robust feature maps, this paper builds a multi-scale feature pyramid. Inspired by FPN [17] and RetinaNet [19], we take the second to fifth convolutional residual blocks to extract feature maps and build our deeper feature pyramid based on them. Generally, while high level feature maps have much semantic information that is beneficial to larger object detection, low level feature maps have abundant detail information that is favorable to small object detection. Thus, we conduct upsampling from higher level feature map to enhance lower level feature map with context information.

In this paper, our feature pyramids are combined with six feature maps, where each feature map has a different scale. Different from RetinaNet [19], our multi-scale feature maps are defined as {P2, P3, P4, P5, P6, P7}, where the strides of them are {4, 8, 16, 32, 64, 128}, respectively. Considering the fact that low level feature maps lack semantic information, we enhance low level feature maps with semantic features upsampled from high level feature maps. Specifically, half of the features of P2 are learned from second convolution block by bottom-up pathway and half upsampled from P3 by top-down pathway, so as P3 and P4. P5 is extracted from fifth convolution block with convolutional operation, while P6 is downsampled from fifth convolution block by max pooling. To obtain additional context information, we further introduce a higher level feature map P7. P7 in our method is downsampled from P6 by 3×3 max pooling with stride 2. Comparatively, in RetinaNet [19], P6 is obtained via a 3×3 stride-2

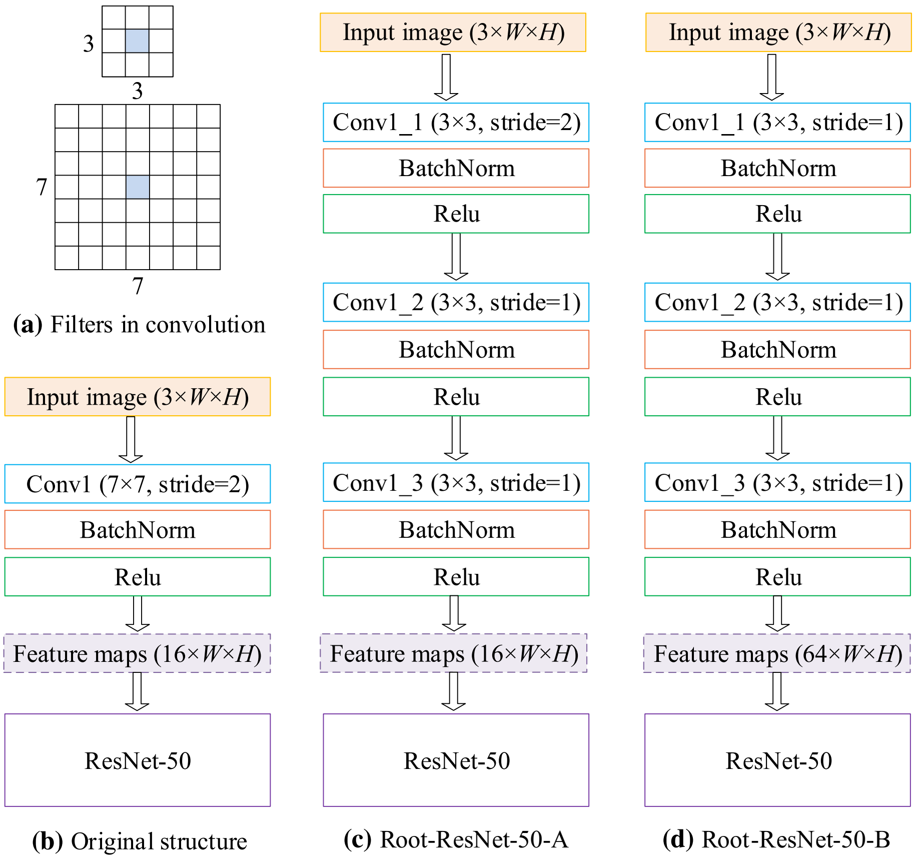


Fig. 3 The architecture of backbone subnetwork. **a** presents filters adopted in original ResNet-50 and our Root-ResNet. **b** is the original structure of ResNet-50, which is beginning from a 7 × 7 convolution layer with stride 2. **c** Root-ResNet-50-A replaces the 7 × 7 convolution layer with three-stacked 3 × 3 convolution blocks. **d** Root-ResNet-50-B further changes the stride size of first convolution layer from 2 to 1

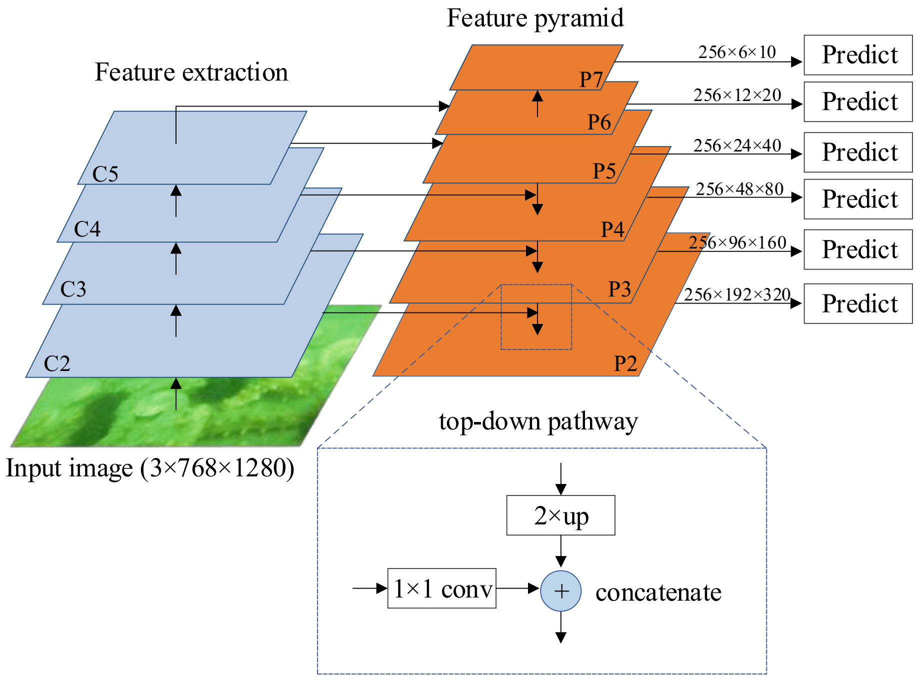


Fig. 4 The representation of proposed multi-scale feature pyramid structure. C2–C5 are behalf of the convolutional blocks. P2–P7 present our feature maps used for the final prediction, where each one is enhanced with features from the bottom-up pathway via lateral connections

conv on C5, and P7 is computed by applying ReLU followed by a 3 × 3 stride-2 conv on P6. The architecture of our feature pyramid is represented in Fig. 4.

The design of our multi-scale feature pyramid network has two main superiorities. On the one hand, it extracts abundant fine-grained information of low level blocks,

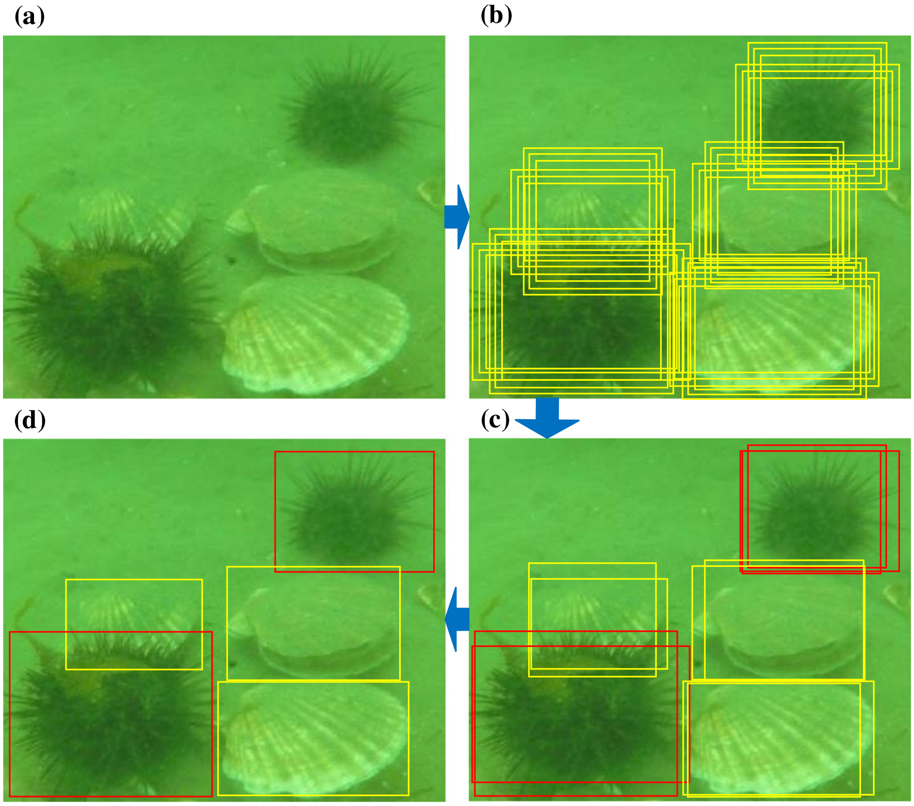


Fig. 5 The detection of proposal. **a** is original image. In **b**, series of region bounding boxes with classification scores and regression offsets are generated from feature maps. It then gets rid of low-score boxes and selects top K proposals in **c**. Finally, NMS method is adopted to remove duplicates with threshold T_t and the detection results are shown in **d**

which are especially beneficial to small object detection. On the other hand, multi-scale feature maps bring richer semantic information that is in favor of large object detection. What is more, it has powerful competence of feature extraction and could locate different scale object with different scale feature map.

3.4 Soft non-maximum suppression for reduplicative box elimination

Reduplicative box elimination plays an essential role in object detection task. Non-maximum suppression (NMS) is the classical suppression method, which could select the bounding box with the maximum detection score for the object and suppress its neighboring boxes. However, traditional NMS has miss suppression issue that may suppress the boxes for its neighboring targets at the same time. To settle this issue, this paper replaces NMS with Soft-NMS to eliminate the duplicates.

In this paper, region proposal network is adopted to generate bounding boxes. For each pixel in images, we take nine anchors, three different scales of width and height, to produce bounding boxes. Each box is classified as a specific label with scores. Considering the fact that these boxes include both valid and invalid detections, we sort the boxes based on the score and select top K ($K = 2000$) detections as proposals. What is more, the redundant proposals can be eliminated by suppression method.

As illustrated in Fig. 5, all of the detection boxes in (b) are sorted based on their scores and the detection box B with the maximum score is selected as the proposal. The intersection of union (IoU) between B and other box B_i is calculated as follows:

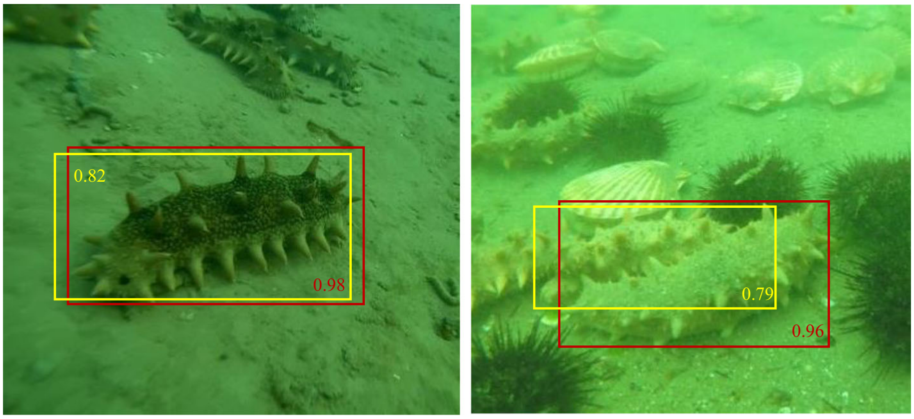
$$\mathrm{IoU} = \frac{B \cap B_i}{B \cup B_i}. \tag{1}$$

Meanwhile, the other detection boxes with a valid overlap with B, which is according to a predefined threshold T_t ($T_t = 0.5$), are suppressed. The process is recursively applied on the remaining boxes.

For traditional NMS algorithm in current object detection pipeline, it is bothered by the miss elimination. After revisiting the NMS method in greater detail, the suppressing process in the NMS algorithm can be described as follows:

$$S_i = \begin{cases} S_i, & \mathrm{IoU}(B, B_i) < T_t \\ 0, & \mathrm{IoU}(B, B_i) \geq T_t \end{cases}, \tag{2}$$

where S_i is the score of box and T_t means the predefined overlap threshold between detected box and true bounding box. As shown in Fig. 6, while the red box is selected as proposal, the yellow box that has a significant overlap with



(a) **Correct suppression** **(b)** **Miss suppression**

Fig. 6 The representation of duplicate elimination. There are two detected bounding boxes with different scores in each images. The overlap of these two boxes lies within the predefined threshold. NMS method sets the score of yellow box to 0.0; this means it will be eliminated. For **b**, this process may suppress the candidate of the adjacent object and increase the miss-rate. Soft-NMS believes it is better to allot a lower score to yellow box, because it has the chance to be selected for the next object (color figure online)

red box will be suppressed by resetting the score as 0. However, in Fig. 6b, because of the miss suppression, the yellow box has no chance to participate in the following selection. That may lead to the reduction of accuracy for the scallop detection. What is worse, this situation exists in the whole detection and is severe for dense marine object detection task.

To settle the issue, this paper adopts Soft-NMS method to suppress the duplicates. Different from NMS, Soft-NMS method resets the sore of yellow box as a lower one. The process of Soft-NMS can be formulated as follows,

$$S_i = \begin{cases} S_i, & \mathrm{IoU}(B, B_i) < T_t \\ S_i(1 - \mathrm{IoU}(B, B_i)), & \mathrm{IoU}(B, B_i) \geq T_t \end{cases}. \tag{3}$$

As a result, the yellow boxes will have the opportunity to be selected as the proposal for the adjacent objects.

Comparatively speaking, Soft-NMS method tactfully resets the scores of duplicates from 0 to a low but nonzero value. So these boxes could participate the following selection. It is critical to avoid miss elimination in marine object detection. Because of the aggregation effect, marine objects in the captured underwater images are usually dense. So miss suppression issue widely exists in marine object detection task. Soft-NMS could overcome this defect in reduplicative boxes removal. The experiments conducted in the next section also validate the effectiveness of Soft-NMS in marine object detection.

3.5 Loss function

In this paper, our training loss function for an image is defined as:

$$L(\{p_i\}, \{t_i\}) = \frac{1}{N_{\mathrm{cls}}} \sum_i L_{\mathrm{cls}}\left(p_i, p_i^*\right) + \lambda \frac{1}{N_{\mathrm{reg}}} \sum_i \mathbb{1}_{\{p_i^* > 0\}} L_{\mathrm{reg}}\left(t_i, t_i^*\right), \tag{4}$$

where i is the index of an anchor in a mini-batch. p_i denotes the predicted probability of anchor i being an object and p_i^* is behalf of the ground-truth label, which is 1 if the anchor is positive and is 0 if the anchor is negative. The predicted bounding box is represented as a vector t_i, which is combined with 4 parameterized coordinates. Meanwhile, the ground-truth box associated with a positive anchor is denoted as a vector t_i^*. The classification loss L_{cls} and the regression loss L_{reg} are set as in [5]. $\mathbb{1}_{\{p_i^* > 0\}}$ represents the indicator function, being 1 if $p_i^* > 0$ and 0 otherwise. This term controls that the regression loss is activated only for positive anchors ($p_i^* = 1$) and is disabled otherwise ($p_i^* = 0$). To make sure the cls term and reg term in Eq. (4) in same dimension, the cls term is normalized by mini-batch size ($L_{\mathrm{cls}} = 256$) and the reg term is normalized by the number of anchor locations ($L_{\mathrm{reg}} = 2400$). λ is the balance weight for L_{reg}, which has been tested in [5] that the detection results are insensitive to λ in a wide range from 1 to 100. Thus, we set $\lambda = 10$ to balance the weight in this paper, which makes both cls and reg terms roughly equally weighted after normalization.

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 the 3 category underwater image datasets and the 20 category PASCAL

VOC datasets [1], respectively. First of all, we execute experiments on underwater image datasets to solve marine object detection task and research the effectiveness of each component in our algorithm. Then, we perform experiments on the PASCAL VOC datasets to compare with the common practice in both accuracy and speed and analyze the performance of our method. The experimental results reveal that our proposed method performs well not only on underwater image datasets, but also on standard datasets. In addition, we also conduct experiments on the 10 category smart unmanned vending machines (UVMs) datasets [45, 46] to test the generalization ability of our method. Illustratively, this paper adopts mean average precision (mAP) as evaluation criterion of accuracy and frames per second (FPS) to test the speed of detection.

4.1 Training details

We take ResNet-50 as our backbone networks, and the base ResNet-50 model is pretrained on ImageNet1k classification set [47]. Unless specified, our network is trained with stochastic gradient descent (SGD) for 100K iterations with the initial learning rate of 0.001, which is reduced by a factor of 10 at iteration 60K and 80K, respectively. We use a weight decay of 0.0001 and a momentum of 0.9. In addition, the input images are resized to 1280 × 768. All of the experimental results are implemented using a Nvidia GeForce GTX 1080 Ti GPU and cuDNN v5.1 and an Intel Core i7-6700K@4.00 GHz.

4.2 Experiments on underwater image datasets

The underwater image datasets are built with the same layout of PASCAL VOC datasets, which mainly include 25,400 pictures with three categories: sea cucumber, sea urchin, and scallop. In order to actually research the detection of marine objects, we capture underwater images with our integrated underwater robot in naturalistic ocean environment and label them by ourselves. To improve the multiplicity of the datasets, we augment the datasets by doing mirror transformation and image enhancement for some pictures.

We represent some instances of the underwater image datasets in Fig. 7. Apparently, the underwater images are blur and color cast. And the scales of marine objects in underwater images are small. What is more, some marine objects, such as sea cucumbers and scallops, have protective coloration to hide themselves into surroundings. Because of the living habits of marine objects, the captured images are usually have a high density of targets. These natures aggravate the challenges of marine object detection task. In accordance with the proposed algorithm, we perform series of experiments on underwater image datasets.

4.2.1 Comparison with popular detectors

We conduct experiments on underwater image datasets with different popular detectors. Specifically, we reimplement popular detectors with default setting on underwater image datasets. The comparison results are shown in Table 1. Apparently, the detection performance on marine objects cannot catch the one on common classes. For

Table 1 Comparison with popular detectors on the underwater image datasets

Approach	Backbone	Input size	FPS	mAP (%)
Fast R-CNN	VGGNet	∼ 1000 × 600	0.4	63.77
Faster R-CNN	ZFNet	∼ 1000 × 600	14	61.95
Faster R-CNN	VGGNet	∼ 1000 × 600	5.6	69.16
Faster R-CNN	ResNet-101	∼ 1000 × 600	3.4	71.01
YOLO	GoogLeNet	448 × 448	41	61.18
YOLOv2	Darknet-19	416 × 416	**61**	73.86
YOLOv3	Darknet-53	416 × 416	30	74.43
SSD	VGGNet	300 × 300	42	70.03
FPN	ResNet-50	∼ 1280 × 768	4.1	74.25
SA-FPN(ours)	ResNet-50	∼ 1280 × 768	3.5	**76.27**

Bold values indicate the best results

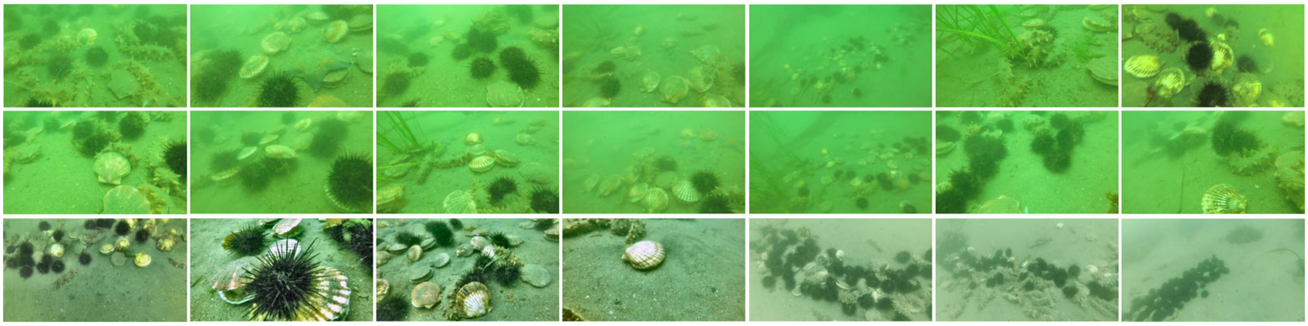


Fig. 7 The images in underwater image datasets. The first row and second row show original underwater images. The last row shows the images after enhancement processing. We augment our datasets by enhancing some underwater images

instance, mAP of Fast R-CNN on underwater image datasets is 63.77%, where 70.0% mAP is achieved on the PASCAL VOC datasets (shown in Table 5). It is because underwater images are complicated and scales of marine objects are usually small.

For object detection task, Faster R-CNN and YOLO are classical approaches. And several improving versions are proposed in recent years. Thus, this paper reimplements different versions of these methods on underwater image datasets. As represented in Table 1, Faster R-CNN with ZFNet just achieves 61.95% mAP. While adopting complicated structure, Faster R-CNN with VGGNet and ResNet-101 could obtain 69.16% and 71.01%, respectively. Comparatively, YOLO series methods have superiority on detection speed. YOLO detector realizes a very fast detection, which could process 41 frames within one second. What is more, YOLOv2 could process 61 frames per second at the mAP of 73.86%. YOLOv3 further improves the detection accuracy on underwater image datasets from 73.86% to 74.43. In addition, SSD detector obtains feature pyramid networks gets 74.25%. Our proposed method performs best on the underwater image datasets with 76.27% mAP, and we will detailedly analyze the effectiveness of our algorithm in the following.

4.2.2 Ablation study

To verify our design of proposed algorithm, we conduct series of ablation experiments to show the comparative effect of each component. In Table 2, we execute FPN on underwater image dataset as baseline and introduce our design on it to improve the performance.

Specifically, the results between first two rows demonstrate that after introducing the high level feature pyramid, the performance on marine object detection is improved. That is benefited with the richer semantic information generated by high level feature pyramid. What is more, the comparison of first row with third row illustrates that the redesign of backbone network also has contributions on detection performance, because it could extract more abundant feature than original structure.

By contrast with FPN, the algorithm proposed in this paper has advantage on marine object detection. Especially, via replacing non-maximum suppression method with Soft-NMS, our method could avoid miss elimination in duplicate removal. With the same setting of experiments, we outperform FPN by 2.02%.

In addition, in terms of these three category objects, the detection of the scallop is performed well than the others, and the sea urchin is much harder to recognize under the water.

4.2.3 Research on backbone subnetwork

We analyze the structure of ResNet and VGGNet and redesign the backbone subnetwork, called Root-ResNet. Specifically, our model is designed based on the ResNet-50 backbone network in experiments. Each convolution layer in our backbone network is adjacent to the BatchNorm layer and ReLU layer. To explore the effect of backbone subnetwork on detection performance, the experiments on different types of backbone structure are carried out.

In contrast to VGGNet, the original structure in ResNet-50 uses relatively large kernel size 7 × 7 with stride 2. As shown in Table 3, it only produces 74.25% mAP on underwater image datasets. Aiming to explore the effect of the kernel size of the first convolution layer on the detector, we attempt several experiments. As illustrated in the first three rows of Table 3, while reducing the scale of filters in first convolutional layer, the performance on detection has been improved slightly. By replacing the kernel size of first convolutional layer from 7 × 7 to 3 × 3, 0.33% mAP is gained.

Activated by DSOD, we decide to replace the first 7 × 7 convolution layer with several 3 × 3 convolution layers. After introducing the stacked convolution layers, we found that the speed is slower than original structure. To study the impact of number of stacked convolution layers in the backbone subnetwork, a group experiments are conducted and the results are shown in Table 3. As the number of convolution layers increases from 1 to 3, the detection results are improved from 74.58 to 74.81%. Considering the cost of computation, we take three 3 × 3 convolution layers as the basic structure in our backbone subnetwork. In addition, we also test the effect on stride size in three-

Table 2 Ablation experiments on underwater image dataset

Network	AP (%)			mAP (%)
	Sea cucumber	Sea urchin	Scallop	
FPN	71.20	70.92	80.64	74.25
FPN + P7	72.34	71.78	81.36	75.16
FPN + root	71.87	71.59	80.98	74.81
FPN + root + P7	72.61	72.32	81.66	75.53
FPN + root + P7 + Soft-NMS(ours)	73.25	73.09	82.47	76.27

Table 3 The exploration on how the structure of backbone network affects the performance on detection

Backbone subnetwork	AP (%)			mAP (%)
	Sea cucumber	Sea urchin	Scallop	
7 × 7, stride = 2	71.20	70.92	80.64	74.25
5 × 5, stride = 2	71.39	71.12	80.75	74.42
3 × 3, stride = 2	71.53	71.31	80.90	74.58
3 × 3, stride = 2	71.69	71.45	80.93	74.69
3 × 3, stride = 1				
3 × 3, stride = 2	71.87	71.59	80.98	74.81
3 × 3, stride = 1				
3 × 3, stride = 1				
3 × 3, stride = 1	**72.34**	**71.78**	**81.36**	**75.16**
3 × 3, stride = 1				
3 × 3, stride = 1				

Bold values indicate the best results

stacked convolution block and get 75.16% mAP on detection.

4.2.4 Effectiveness of learning rate

To explore the effect of learning rate in training, we design several experiments with different learning rates. With the same setting in other components in our method, we only adjust the learning rate on training phase and observe the detection results.

As demonstrated in Table 4, with the reduction of learning rate, the performance on detection has been improved. While the learning rate is set as 0.001, our model gets 76.27% mAP on underwater image datasets. However, too lower learning rate will affect the rate of convergence of the network. So, we stop reducing learning rate after 0.001. In this paper, we choose 0.001 as our default learning rate.

4.2.5 Analysis on training

After repeated experiments, SA-FPN improves the performance on marine object detection, even in seafood serried scene. The comparison of precision–recall curve between FPN and SA-FPN is shown in Fig. 8. While FPN achieves 74.25% mAP on underwater image dataset, our method could perform 76.27% mAP.

Table 4 Analysis of learning rate for our proposed method on underwater image dataset

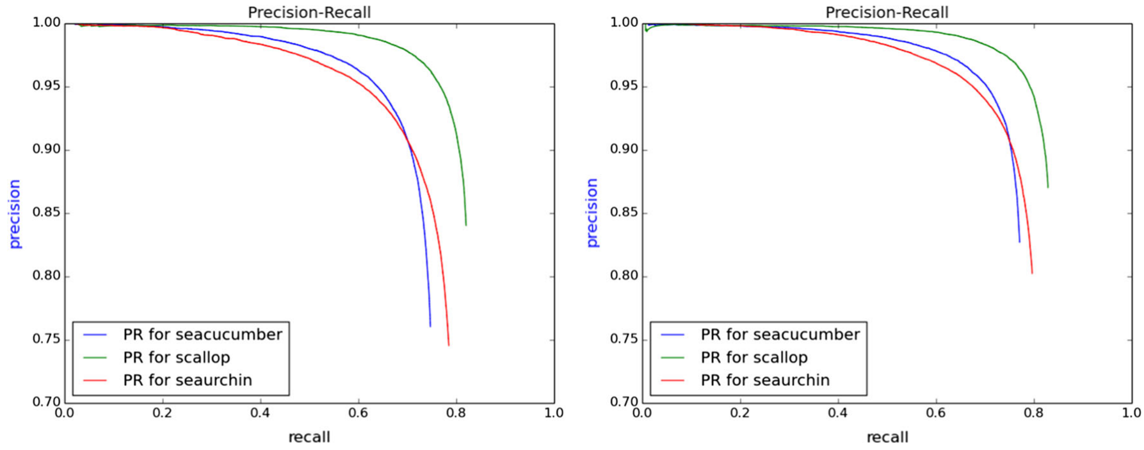
Learning rate	AP (%)			mAP (%)
	Sea cucumber	Sea urchin	Scallop	
0.05	72.99	72.58	82.10	75.89
0.01	73.01	72.60	82.12	75.91
0.005	73.20	72.82	82.37	76.13
0.001	73.25	73.09	82.47	76.27

Furthermore, to analyze the variation of our method with FPN in training, we visualize the loss between these two algorithms. As demonstrated in Fig. 9, our method could converge quickly with lesser amplitude fluctuation of loss in training.

With the intension of explaining the performance on our method, we arbitrarily take some detection results on different methods as examples. In Fig. 10, while first row results are conducted on FPN, second row results on our algorithm. By carefully comparing each group image, we found that our method outperforms FPN on marine object detection. Specifically, SA-FPN could detect much more objects in images, especially for the small one, that may leave out with FPN method.

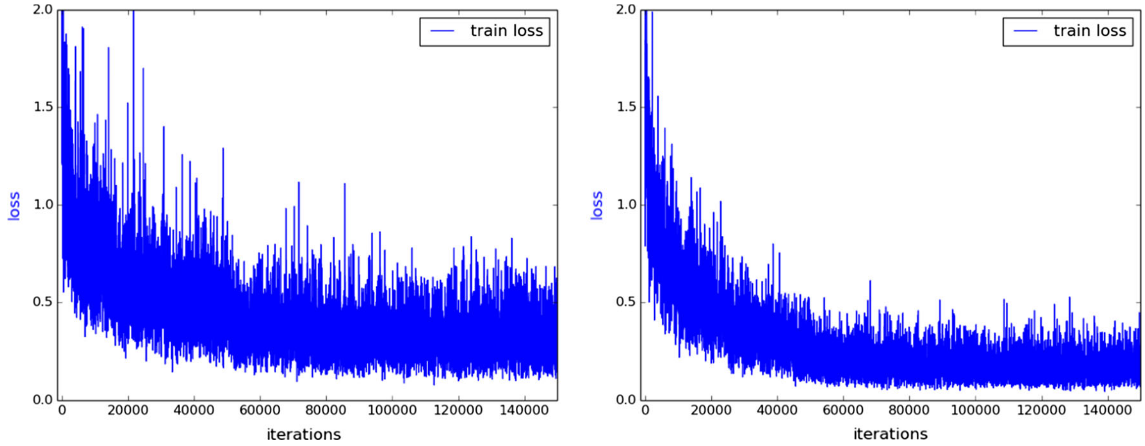
More detection results of our algorithm are shown in Fig. 11. What revealed in Fig. 11 is that SA-FPN performs well in different situations that varied from small serried scene to bigger one. Even in weedy and muddy water environment, our method still could detect the target accurately.

However, our algorithm still faces with challenges on marine object detection task. It is extremely difficult to distinguish very close objects with same category from each other. In addition, detection performance on shadowed objects also needs to be improved. Some failure cases on marine object detection experiments are given in Fig. 12. For example, while sea urchins are very close to each other, it is hard to figure out whether they are regarded as one target or not, so do as sea cucumber and scallop. Besides, when the scallop is almost shadowed by sea urchin in Fig. 12, it would be ignored to detect.



(a) FPN (b) SA-FPN

Fig. 8 The comparison of precision–recall curve between FPN and SA-FPN



(a) FPN (b) SA-FPN

Fig. 9 The comparison of training loss between FPN and SA-FPN. Comparatively, our method is converged quickly with lesser amplitude fluctuation of loss

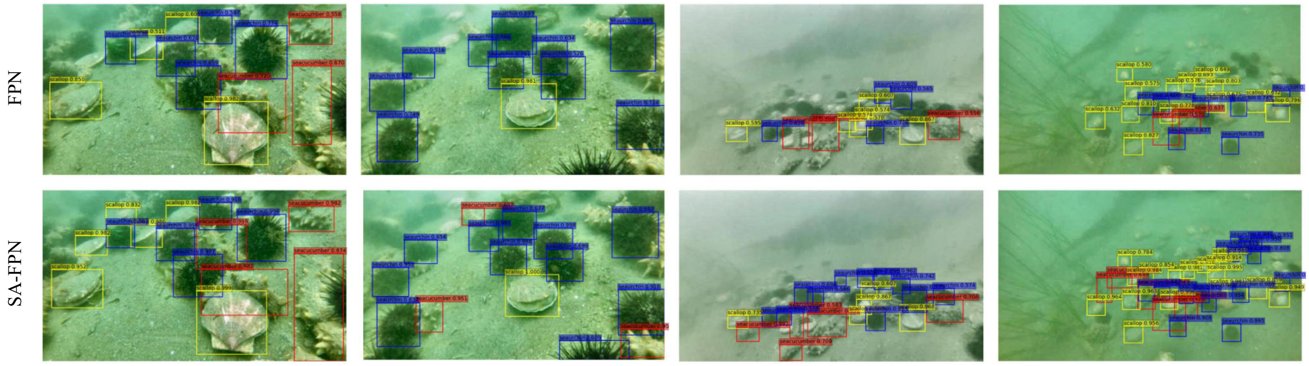


Fig. 10 Qualitative detection results of FPN and SA-FPN algorithm on underwater image dataset. Top row shows detection results on FPN method, and bottom row shows experiments on SA-FPN. By comparison, our method could detect more targets than FPN method in underwater image and improve the performance on marine object detection

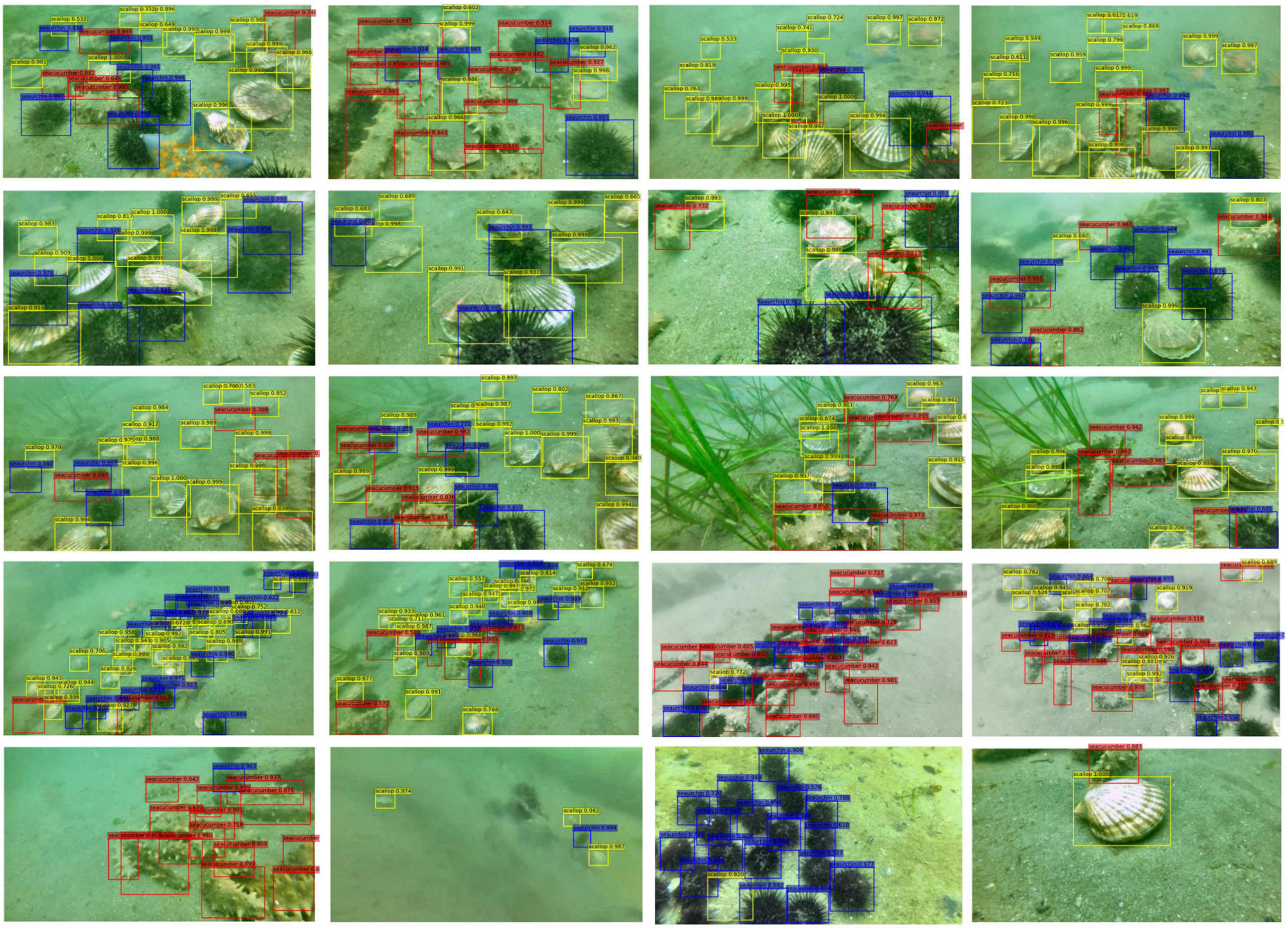
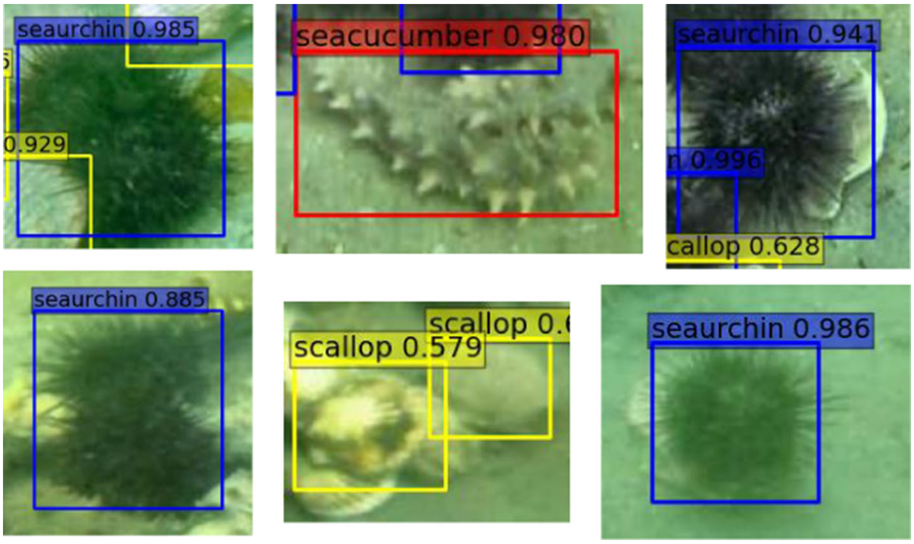


Fig. 11 Some detection results of SA-FPN method on underwater image dataset. Our method works well on marine object detection, even in serried scene

Fig. 12 Failure cases on marine object detection task



4.3 Experiments on Pascal VOC datasets

To verify the effect of our proposed method on standard object detection dataset, we conduct the experiments on the PASCAL VOC datasets. Specifically, we train the model on the VOC 2007 and VOC 2012 trainval sets (16,551 images) and test on the VOC 2007 test set (4952 images). We compare SA-FPN with the state-of-the-art object detection approaches on the PASCAL VOC 2007 datasets in Table 5.

On the basis of whether region proposal is needed or not, approaches of object detection are usually divided into one-stage detectors and two-stage detectors. Two-stage detectors firstly generate region proposal from feature maps and then detect based on these proposals. One-stage detectors frame object detection as a regression issue and take a single neural network to detect object and category from fully images in one evaluation.

As shown in Table 5, one-stage detectors have advantages on detection speed. For example, YOLOv2 could real-timely detect targets with speed of 67 FPS, and SSD300 may reach 46 FPS on detection task. The upgrade vision of these detectors, for instance YOLOv3, SSD512, DSSD321, and GFR-DSOD300, achieve high detection accuracy on the cost of increasing computation burden. GFR-DSOD300 even achieves 78.9% mAP. As always, two-stage detectors get satisfactory detection accuracy. For example, faster R-CNN with ResNet-101 obtains 76.4% mAP and FPN gets 77.1% on the PASCAL VOC 2007 datasets with default setting. R-FCN and MR-CNN improve the detection accuracy to 77.4% and 78.2%, respectively. Finally, we achieve 79.1% mAP and outperform FPN with 2% mAP on PASCAL VOC dataset.

In addition, we investigate the effectiveness of each component of our SA-FPN framework. We design several controlled experiments on the PASCAL VOC 2007 datasets for ablation study. As shown in Table 6, we implement FPN with default setting on the PASCAL VOC 2007 datasets and get 77.06% mAP. Then, we add each component on original FPN and observe the function of it. From the experimental results, we found that the backbone subnetwork carries out 0.39% improvement and our multiscale feature pyramid improves 0.79% mAP. What is more, by combining these components together, our framework achieves 79.13% mAP on the PASCAL VOC datasets.

Table 6 Ablation experiments on the PASCAL VOC 2007 datasets

Network	mAP (%)
FPN [17]	77.06
FPN + P7	77.85
FPN + root	77.45
FPN + root + P7	78.21
FPN + root + P7 + Soft-NMS(ours)	79.13

Table 5 Detection results on the PASCAL VOC 2007 datasets

Approach	Backbone	Input size	FPS	mAP (%)
One-stage detectors				
YOLO [6]	GoogLeNet	448 × 448	45	63.4
YOLOv2 [29]	Darknet-19	416 × 416	67	76.8
YOLOv3 [31]	Darknet-53	416 × 416	34	77.2
RON384 [30]	VGGNet	384 × 384	15	75.4
SSD300 [7]	VGGNet	300 × 300	46	74.3
SSD512 [7]	VGGNet	512 × 512	19	76.8
DSSD321 [16]	ResNet-101	321 × 321	9.5	78.6
DSOD300 [32]	DS/64-192-48-1	300 × 300	17.4	77.7
GFR-DSOD300 [33]	DS/64-192-48-1	300 × 300	17.5	78.9
Two-stage detectors				
Fast R-CNN [4]	VGGNet	∼ 1000 × 600	0.6	70.0
Faster R-CNN [5]	VGGNet	∼ 1000 × 600	7	73.2
Faster R-CNN [24]	ResNet-101	∼ 1000 × 600	5	76.4
ION [26]	VGGNet	∼ 1000 × 600	1.25	75.6
FPN [17]	ResNet-50	∼1280 × 768	5	77.1
R-FCN [27]	ResNet-50	∼ 1000 × 600	11	77.4
MR-CNN [25]	VGGNet	∼ 1000 × 600	0.03	78.2
SA-FPN(ours)	ResNet-50	∼1280 × 768	4	**79.1**

Bold value indicates the best result

Table 7 Detection results on the smart UVMs datasets

	YOLO-v2	YOLO-v3	Faster R-CNN	R-FCN	SSD	DSSD	SA − FPN(ours)
FT	90.90	90.90	90.91	90.90	90.91	90.90	90.90
NE	90.91	90.91	90.89	90.90	90.90	90.89	90.91
JDB	90.88	90.91	90.91	90.91	90.91	90.90	90.91
MZ	90.91	99.99	90.91	96.37	99.92	90.90	99.96
JGMT	90.84	90.90	90.73	90.88	90.87	90.85	90.88
GTEN	90.82	90.91	90.62	90.88	90.88	90.83	90.91
UAMT	90.90	90.89	90.83	90.86	90.86	90.86	90.90
VVM	90.77	90.87	90.78	90.85	90.85	90.85	90.88
IBT	90.89	90.91	90.91	90.90	90.90	90.89	90.90
MM	90.90	90.88	90.88	90.91	90.89	90.89	90.91
mAP (%)	90.87	**91.81**	90.84	91.43	91.79	90.88	**91.81**

Bold values indicate the best results

4.4 Experiments on smart UVMs datasets

In order to test the generalization ability of our algorithm, we also train and evaluate our method on the smart UVMs datasets. Smart UVMs datasets are compiled for object detection in unmanned retail application environments, which contain over 30,000 images captured in a refrigerator equipped with different cameras [45, 46]. For the static detection task in Smart UVMs datasets, there are 34,052 images with 10 kinds of beverages in the dataset, including 14,651 images in the training set, 14,040 images in the validation set, and 5361 images in the testing set.

We train our method on smart UVMs datasets and compare with several state-of-the-art object detection models in Table 7. For representation, label of fenda is changed to FT, yingyangkuaixian to NE, jiaduobao to JDB,

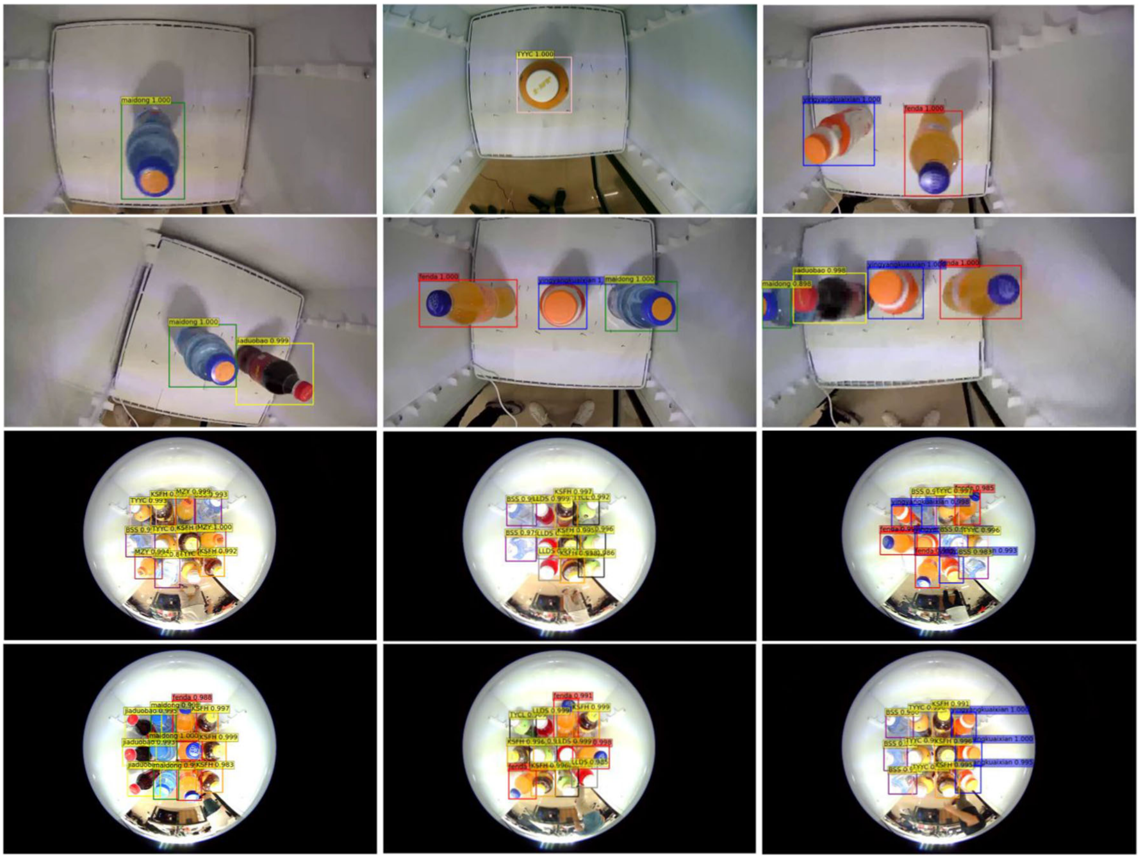


Fig. 13 Qualitative detection results on smart UVMs datasets

maidong to MZ, TYCL to JGMT, BSS to GTEN, TYYC to UAMT, LLDS to VVW, KSFH to IBT, and MZY to MM [45].

As shown in Table 7, our method could reach the best performance of 91.81% mAP, the same as YOLOv3, on detection task with smart UVMs datasets. In addition, qualitative detection results of our model on smart UVMs datasets are shown in Fig. 13. Our method could detect targets accurately.

5 Conclusion

This paper proposes a scale-aware feature pyramid network to detect marine objects. Firstly, we propose a special backbone subnetwork architecture called Root-ResNet on the foundation of ResNet-50 to extract fine-grained feature maps. Root-ResNet improves 0.36% mAP on marine object detection task by replacing the first 7 × 7 convolution layer with three-stacked 3 × 3 convolution blocks. What is more, we build a multi-scale feature pyramid to enhance the semantic features. 0.91% mAP and 0.79% mAP are gained on underwater image datasets and PASCAL VOC 2007 datasets, respectively. Finally, to suppress the reduplicative bounding boxes on the targets, this paper adopts soft non-maximum suppression algorithm to replace NMS, which may cause miss elimination. The experimental results reveal that our methods have effective performance on marine object detection. After several experimental tests, our methods could reach 76.27% mAP on marine object detection and outperform FPN by 2.02%. In addition, we also train and evaluate on the smart UVMs datasets to test the generalization ability of our algorithm and achieve the best performance of 91.81% mAP.

In the future, we will continue to exploit potentialities of convolutional neural network on marine object detection and improve the performance on turbid and crowded environment.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China Grant 61370142 and Grant 61272368, by the Fundamental Research Funds for the Central Universities Grant 3132016352, by the Fundamental Research of Ministry of Transport of P. R. China Grant 2015329225300, by Liaoning Revitalization Talents Program, XLYC1908007, by the Dalian Science and Technology Innovation Fund 2018J12GX037, by the Dalian Science and Technology Innovation Fund 2019J11CY001 and Dalian Leading talent Grant, by the Foundation of Liaoning Key Research and Development Program, China Postdoctoral Science Foundation 3620080307.

Compliance with ethical standards

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

References

1. Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. Int J Comput Vis 88(2):303–338
2. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: common objects in context. In: European conference on computer vision, pp 740–755. Springer
3. 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
4. Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 1440–1448
5. 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
6. 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
7. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2016) Ssd: single shot multibox detector. In: European conference on computer vision. Springer, pp 21–37
8. Kashif I, Salam RA, Azam O, Talib AZ (2007) Underwater image enhancement using an integrated colour model. IAENG Int J Comput Sci 34(2):239–244
9. Schettini R, Corchs S (2010) Underwater image processing: state of the art of restoration and image enhancement methods. EURASIP J Adv Signal Process 2010(1):746052
10. Serikawa S, Huimin L (2014) Underwater image dehazing using joint trilateral filter. Comput Electr Eng 40(1):41–50
11. Li C-Y, Guo J-C, Cong R-M, Pang Y-W, Wang B (2016) Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior. IEEE Trans Image Process 25(12):5664–5677
12. Chiang JY, Chen Y-C (2011) Underwater image enhancement by wavelength compensation and dehazing. IEEE Trans Image Process 21(4):1756–1769
13. Cai Z, Fan Q, Feris RS, Vasconcelos N (2016) A unified multi-scale deep convolutional neural network for fast object detection. In: European conference on computer vision. Springer, pp 354–370
14. Zhang H, Wang K, Tian Y, Gou C, Wang F-Y (2018) Mfr-cnn: incorporating multi-scale features and global information for traffic object detection. IEEE Trans Veh Technol 67(9):8019–8030
15. Zheng C, Yang M, Wang C (2017) A real-time face detector based on an end-to-end CNN. In: 2017 10th international symposium on computational intelligence and design (ISCID). IEEE, vol 1, pp 393–397
16. Fu C-Y, Liu W, Ranga A, Tyagi A, Berg AC (2017) DSSD: Deconvolutional single shot detector. arXiv preprint arXiv:1701.06659
17. 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
18. 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
19. Lin T-Y, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp 2980–2988

20. Tian Z, Shen C, Chen H, He T (2019) Fcos: fully convolutional one-stage object detection. arXiv preprint arXiv:1904.01355
21. 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
22. Kirillov A, Girshick R, He K, Dollár P (2019) Panoptic feature pyramid networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6399–6408
23. Pang J, Chen K, Shi J, Feng H, Ouyang W, Lin D (2019) Libra r-cnn: towards balanced learning for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 821–830
24. 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
25. 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
26. Bell S, Zitnick CL, Bala K, Girshick R (2016) Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2874–2883
27. 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
28. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 2961–2969
29. Redmon J, Farhadi A (2017) Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7263–7271
30. Kong T, Sun F, Yao A, Liu H, Lu M, Chen Y (2017) Ron: reverse connection with objectness prior networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5936–5944
31. Redmon J, Farhadi A (2018) Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767
32. 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
33. 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
34. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3431–3440
35. Hariharan B, Arbeláez P, Girshick R, Malik J (2015) Hypercolumns for object segmentation and fine-grained localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 447–456
36. Liu W, Rabinovich A, Berg AC (2015) Parsenet: looking wider to see better. arXiv preprint arXiv:1506.04579
37. Kong T, Yao A, Chen Y, Sun F (2016) Hypernet: towards accurate region proposal generation and joint object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 845–853
38. Rothe R, Guillaumin M, Van Gool L (2014) Non-maximum suppression for object detection by passing messages between windows. In: Asian conference on computer vision. Springer, pp 290–306
39. Hosang J, Benenson R, Schiele B (2016) A convnet for non-maximum suppression. In: German conference on pattern recognition. Springer, pp 192–204
40. Hosang J, Benenson R, Schiele B (2017) Learning non-maximum suppression. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4507–4515
41. Bodla N, Singh B, Chellappa R, Davis LS (2017) Soft-NMS–improving object detection with one line of code. In: Proceedings of the IEEE international conference on computer vision, pp 5561–5569
42. Jiang B, Luo R, Mao J, Xiao T, Jiang Y (2018) Acquisition of localization confidence for accurate object detection. In: Proceedings of the European conference on computer vision (ECCV), pp 784–799
43. He Y, Zhu C, Wang J, Savvides M, Zhang X (2019) Bounding box regression with uncertainty for accurate object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2888–2897
44. Zhu R, Zhang S, Wang X, Wen L, Shi H, Bo L, Mei T (2019) Scratchdet: training single-shot object detectors from scratch. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2268–2277
45. Zhang H, Li D, Ji Y, Zhou H, Wu W (2019) Deep learning-based beverage recognition for unmanned vending machines: an empirical study. In: 2019 IEEE 17th international conference on industrial informatics (INDIN). IEEE, vol 1, pp 1464–1467
46. Zhang H, Li D, Ji Y, Zhou H, Liu K (2019) Towards new retail: a benchmark dataset for smart unmanned vending machines. IEEE Trans Ind Inform PP(99):1
47. 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 Vis 115(3):211–252