ISSN: 2664-9640

Research on marine organism detection algorithm based on improved YOLOv5

Chao Xu [a], Xianjiu Guo [b,*], Ziwen Chen [c] and Ting Liu [d]

Dalian Ocean University, Dalian 116023, China;

[a] 15827610565@163.com, [b] gxj@dlou.edu.cn, [c] chenziwen@dlou.edu.cn ,

[d] 1174622725@qq.com;

Corresponding Author: Xianjiu Guo

Abstract

In order to better grasp the distribution of underwater marine organism, this paper combines YOLOv5s and ShuffleNetV2 to propose a more lightweight object detection model SL-YOLO for marine organism. Firstly, ShuffleNetV2 module is used to replace the C3 module of feature extraction in YOLOv5s. In order to further improve the detection accuracy, BiFPN module is used for feature fusion. Finally, DIOU-NMS is used to replace NMS to solve the missing detection. The results show that compared with the original YOLOv5s model, SL-YOLO is equivalent to the original YOLOv5s model in the mAP, which reduces the number of parameters by 39 %, and achieves the balance between accuracy and real-time performance.

Keywords

Object detection, Detection of marine organism, YOLOv5s, ShuffleNetV2, BiFPN.

1. Introduction

Ocean organism is an important property in marine fishery. It is not only a precious food material, but also a precious medicinal material. The traditional ocean organism detection method is mainly carried out by professional personnel to determine the category and locate the target of the image by extracting the characteristics of the image [1-2]. This method requires not only professional knowledge but also long time and low detection accuracy.

Compared with the traditional machine vision method, the deep learning method can automatically learn the characteristics of the image, save manpower cost and time cost. For example, Han et al. [3] proposed an improved Faster RCNN method to improve the detection accuracy of marine organisms, but the use of two-stage detection algorithm makes the overall detection time cost is relatively high. Zhao Dean et al. [4] proposed a live crab detection method based on improved YOLOv3, but the parameters of the model are large and difficult to run on mobile devices. Yu et al. [5] applied the MobileNet-SSD network model to the detection of ocean organism. This model can significantly improve the detection speed without losing the accuracy, but the simple one-way fusion of different levels of features does not improve the detection accuracy. In this study, an improved object detection model SL-YOLO(A lightweight network based on YOLOv5s and ShuffleNetV2) with high precision and low parameter number of YOLOv5 is proposed :

The feature extraction network of YOLOv5 is replaced by ShuffleNetV2 [6], which reduces the number of parameters and computation of the model and makes the model better suitable for lightweight applications.

ISSN: 2664-9640

It is proposed to replace the feature fusion module PANet [7] of YOLOv5 with BiFPN [8]. This module can make the fusion module have different weight parameters that can be learned, and can better fuse the features at different levels.

According to the situation of multiple overlaps and target occlusion between targets, DIOU-NMS [9] is proposed to replace NMS, so that the adjacent ocean organism treasures can be more detected.

2. Text Construction of SL-YOLO Model

2.1. The algorithm of YOLOv5s

YOLOv5 is a one-stage target detection algorithm [10-14] proposed by Ultralytics LLC. Compared with Two Stages, such as R-CNN [15], Fast-RCNN [16], Faster-RCNN [17], R-FCN [18], FPN [19] etc., the one-stage target detection algorithm does not need to generate Region proposal, and directly generates the category probability and location information of the target. Therefore, the characteristics of the one-stage algorithm are the fast detection speed, which means that the detection accuracy needs to be sacrificed.

YOLOv5 can be divided into s, m, l and x according to the size of the model. Up to now, the latest version is 6.1. Since our target is a target detection algorithm with high detection accuracy but small model, we choose YOLOv5s 6.1 as the reference model of this study.

As shown in Figure 1, YOLOv5s consists of four parts: input, backbone, neck and detection. In the input, Mosaic data enhancement, adaptive anchor frame calculation and adaptive image scaling are used to preprocess the data. The focus layer in the previous version is deleted by backbone, and SiLU is used as the activation function. The C3 structure composed of BottleNeck is stacked to extract the feature of the network. Finally, SPPF is used to replace the original SPP structure for feature fusion. The neck layer uses PANet characteristic pyramid structure to achieve multi-scale fusion; detection uses GIOU _ Loss as the loss function of boundary anchor frame, and uses weighted NMS to screen multiple anchor frames to improve the accuracy of target detection. Finally, three detection heads 76×76、38×38、19×19 are output, and correspond to the detection results of small target, medium target and large target, respectively.

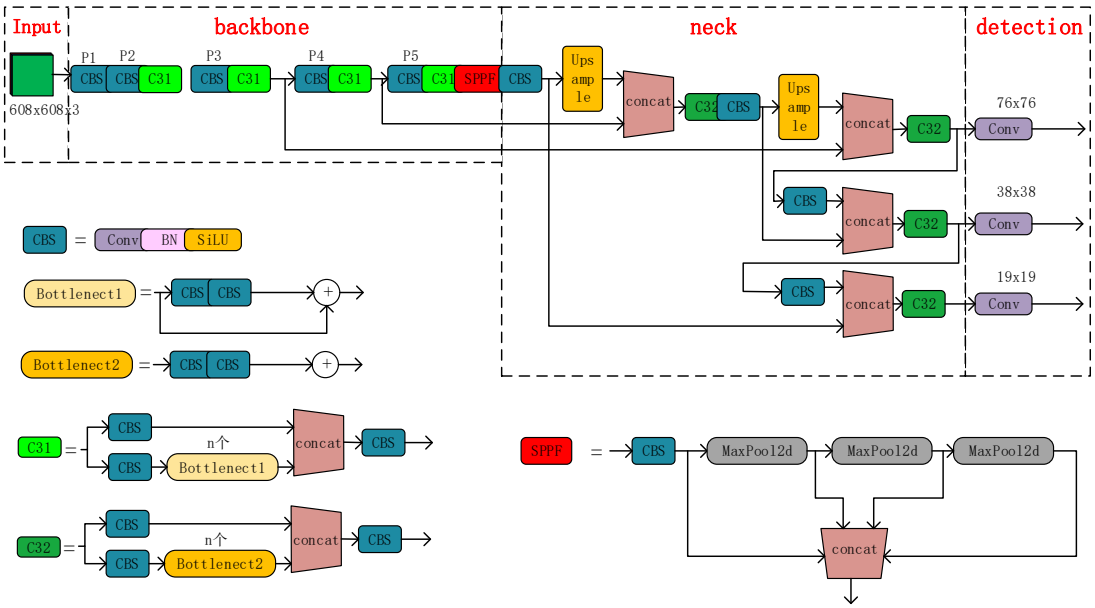

Figure 1: The Structure of YOLOv5s

ISSN: 2664-9640

2.2. The algorithm of ShuffleNetV2

In order to make the model more easily deployed to some small mobile devices, ShuffleNetV2 is used to replace the feature extraction network CSPDarkNet53 in YOLOv5 to ShuffleNetV2. ShuffleNetV2 is a lightweight network model. Based on ShuffleNetV1 [20] , the author of this model proposes four guidelines for the design of efficient lightweight network models :

When the number of channels of the input feature matrix and the output feature matrix of the convolution layer is equal, the memory access cost of the model is the smallest.

When the number of packet convolution increases, MAC will also increase.

The higher the degree of fragmentation of network design, the slower the speed.

The impact of element-level operation cannot be ignored.

Based on these four design criteria, the authors propose two structural units in Fig. 2. Taking the left structure (S1) in Fig. 2 as an example, the left branch does not do any processing to meet the criterion 3), and the right branch cancels the grouping convolution in ShuffleNetV2 to meet the criterion 2). The deep separable convolution is used to reduce the number of parameters of the model, and the number of input and output channels of the three convolution layers is consistent, meeting the criterion 1). Finally, the original element addition operation is replaced by channel splicing, meeting the criterion 4). This structure can not only integrate the information between channels, but also reduce the MAC of the model. The right structure(S2) in Fig. 2 also meets the above four criteria. Since there is no channel segmentation and the step is 2, the channel of the output feature matrix is doubled and the size is halved. The operation method is basically the same as that of S1, which is not repeated here.

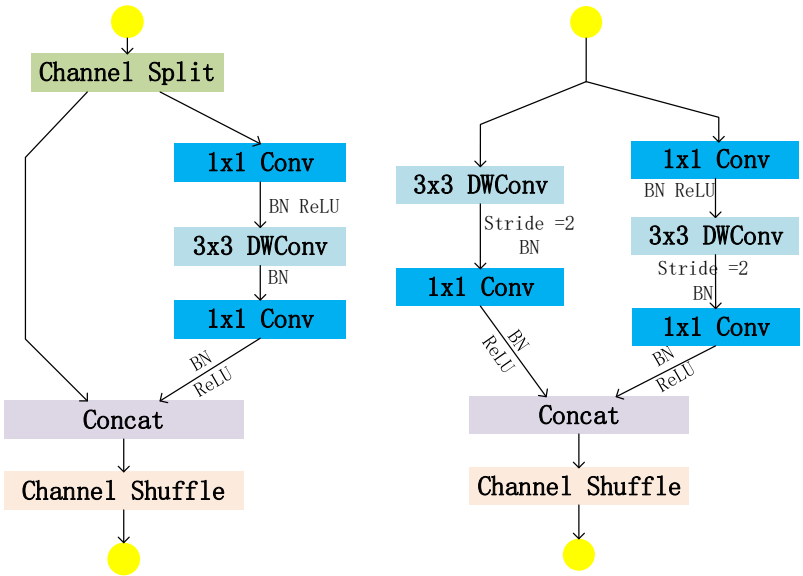

Figure 2: The Structure of ShuffleNetV2

2.3. The Improved YOLOv5s feature extraction network

The feature extraction part of SL-YOLO algorithm proposed in this paper is mainly through stacking the two structures proposed by ShuffleNetV2. For example, S1 is used to replace the C31 module in Figure 1 to extract the feature of the image, and S2 is used to replace the CBS module in Figure 1 to complete the downsampling operation.

Table 1: The composition of improved YOLOv5s feature extraction network

From	Num	Module	Arguments	out
-1	1	CBS	[64,6,2]	[304,64]
-1	1	S2	[128,2]	[152,128]

ISSN: 2664-9640

-1	3	S1	[128]	[152,128]
-1	1	S2	[256,2]	[76,256]
-1	6	S1	[256]	[76,256]
-1	1	S2	[512,2]	[38,512]
-1	9	S1	[512]	[38,512]
-1	1	S2	[1024,2]	[19,1024]
-1	3	S1	[1024]	[19,1024]
-1	1	SPPF	[1024,5]	[19,1024]

2.4. The Improved YOLOv5s Feature Fusion Module

The proposal of ResNet [21] enables us to deepen the level of the network to improve our detection accuracy. High-dimensional features have stronger semantic information but lack sufficient location information, and it is difficult to identify small targets. Therefore, the feature pyramid (FPN) combines the bottom-up high-dimensional features with the adjacent top-down low-dimensional features. PANet (Path Aggregation Network) proposes that adding a bottom-up network behind the FPN to form a two-way fusion can better integrate features and have better performance in target detection. Therefore, both YOLOv4 and YOLOv5 adopt the PANet method as the feature fusion module of the neck part. However, the input features have different resolutions but only the indiscriminate fusion cannot learn effective fusion information.

BiFPN is a weighted bidirectional feature pyramid network proposed on the basis of PANet, which introduces learning weight parameters to reflect the importance of different input features. As shown in the diagram, it first deletes a node in PANet that has only one input edge and no feature fusion, and it has little contribution to the fusion network of different features; then a connection path is added between the input and output nodes of the same scale to integrate more features without introducing more computation. Finally, top-down and bottom-up paths are regarded as a feature module to achieve higher level feature fusion by repeated stacking.

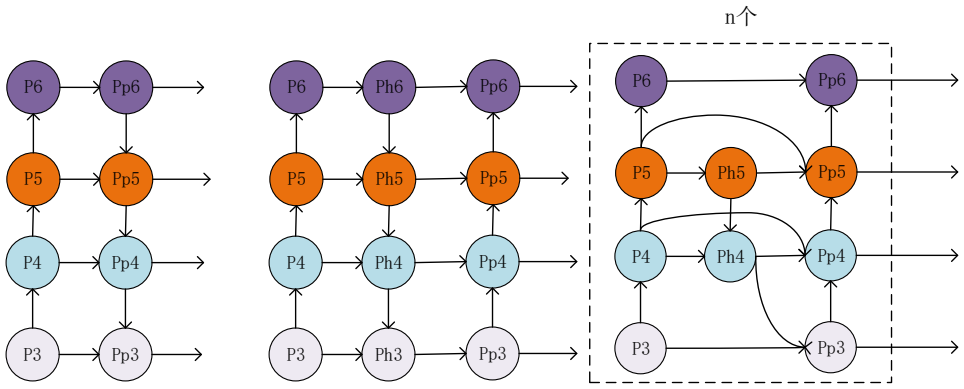

Figure 3: Three kinds of structure,FPN(left),PANet(middle),BiFPN(right)

As shown in the figure, P3 - P6 represents the feature extraction layer, Ph3 - Ph6 represents the intermediate layer between the feature extraction layer and the output layer, and Pp3 - Pp6 represents the output layer. BiFPN uses a fast normalized fusion method, the formula as shown in (1):

$$o = \sum_i \frac{\omega_i}{\varepsilon + \sum_j \omega_j} . I_i \quad (1)$$

ISSN: 2664-9640

Where o presents the output feature, and a small number ε is represented to ensure that the denominator is not 0 (can be set to 0.0001). ω_i, ω_j are weight parameters of learned feature map where $0 < \omega_i, \omega_j < 1$, I_i is the input feature, and the weight of each training is guaranteed to be greater than 0 by the ReLU activation function. Through training, the most appropriate weight will eventually be found to fuse the characteristics of different levels.

2.5. The Improved YOLOv5 detection head module

When there are multiple targets in the required detection area, there may be overlap between multiple targets. The standard of traditional NMS (non-maximum has always been) algorithm is to calculate the IOU (cross-over ratio) between the current detection frame and the detection frame with the highest score. If the value is greater than the set threshold, the current frame will be considered to be repeated with the detection frame with the highest score, and the current detection frame will be deleted. However, if the detection targets in the detection area are very dense, this method is obviously unreasonable. In the process of ocean organism detection, there are a large number of targets gathering and overlapping. Therefore, we use the DIOU-NMS method to select the detection box. The formula of DIOU-NMS is:

$$S_i = \begin{cases} S_i, IOU - R_{DIOU}(M, B_i) < \varepsilon \\ 0, IOU - R_{DIOU}(M, B_i) \geq \varepsilon \end{cases} \tag{2}$$

$$R_{DIOU} = \frac{\rho^2(b, b_t)}{c^2} \tag{3}$$

S_i represents the score of the current prediction box, where ρ^2 is the Euclidean distance, b, b_t are defined as the center points of the prediction box and the real box respectively, c is the diagonal length of the smallest rectangle containing two boxes, M represents the box with the highest score, and ε is the set threshold.

The key to using DIOU-NMS is that if the IOU between the current frame and the highest score frame is large but the center point between the two is relatively far apart, the use of NMS must be to delete the current frame, but using this algorithm not only compares the IOU but also compares the center distance between the two, so as to determine that the current frame cannot be deleted, so as to improve the missed detection.

3. Analysis of experimental results

3.1. Experimental environment and datasets

The operating system used in this study is Windows 10, the CPU is Intel(R) Core(TM) i7-10750H CPU @ 2.60GHz, and the running memory size is 24.0GB. The GPU model is the NVIDIA GeForce RTX 2080 and the memory size is 11GB. This study is based on the Pytorch deep learning framework, using Python as the programming language, CUDA version 10.2.89, CUDANN version 8.0.5.

This research dataset comes from the official dataset of ChinaMM2018 Underwater Robot Target Grabbing Competition, which was taken by divers in the ocean organism breeding area, the dataset are mainly composed of sea cucumbers, sea urchins, scallops, the format of the image is .jpg, the official provides the corresponding xml annotation file at the same time as providing the image dataset, after removing some of the image quality, mismatch, problematic images and annotation files, The dataset of 6213 images of ocean organism used in this study was formed, and finally the dataset was divided into training verification sets and test sets according to the ratio of 8:2.

3.2. Object detection ablation experiment

This design first uses YOLOv5s to train the original dataset and name it YOLOv5s; Replace the YOLOv5 feature extraction network with ShuffleNetV2 and name this module SF; replace the

ISSN: 2664-9640

neck part of YOLOv5 with BiFPN and name this module BF; replace the NMS of the detection part with DIOU-NMS, the SL-YOLO module proposed by the Group Cost Institute, and name the DIOU-NMS module as DN. In order to verify the optimization effect of each module on YOLoOv5s, an ablation experiment was carried out in this study.

Table 2: Ablation experiment based on YOLOv5s

Algorithm	mAP@0.5/%	GFLOP	params/M	FPS
YOLOv5s	87.57	15.3	6.75	57.8
YOLOv5s+SF	82.63	8.9	4.05	95.gf2
YOLOv5s+SF+BF	85.43	9.5	4.26	79.3
SL-YOLO	86.83	9.5	4.26	79.3

As you can see from the table 2, we used yolov5s with a model size of 6.75MB and a floating-point operation volume (GFLOP) of 15.3; after we replaced the feature extraction network with ShuffleNetV2, the parameter amount was reduced by 40%, and the floating-point operation volume was reduced by 41.8%, mAP@0.5 was reduced by only 5 percentage points; after further replacing the neck part with BiFPN, the parameter amount was increased by only 0.19MB, and the mAP was increased by 3 percentage points; and finally the NMS part of the detection part was replaced by DIOU_NMS, and the mAP was only reduced by 0.7% compared to YOOlOv5s, but the parameter amount was reduced by 2.49M, and the FPS also reached 79.5 frames. Without reducing the detection accuracy, the parameters of the model is reduced by 36.9%, making the model more lightweight.

As shown in the figure 4, the training epoch is set to 120 rounds, and at the epoch is around 100, the models are close to convergence, where the use of YOLov5s can make our mAP the highest, the YOLOv5s_SF lowest, and in the case of the same amount of model parameters, the yolov5s_SFBFDN (SL-YOLO) is slightly higher than the yolov5s_SFBF.

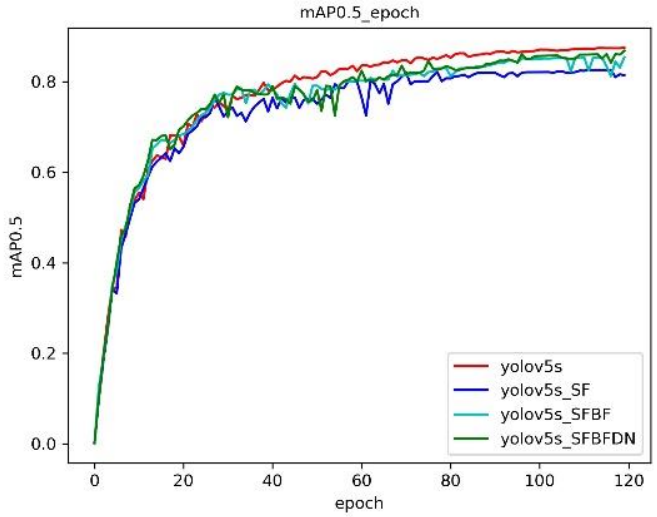

Figure 4: Training mAP graph corresponding to the four models

In order to further prove that we can add the DN module to improve the missed detection situation of our model without changing the number of parameters, 3 sets of pictures are randomly selected for detection, as shown in the following figure 5:

ISSN: 2664-9640

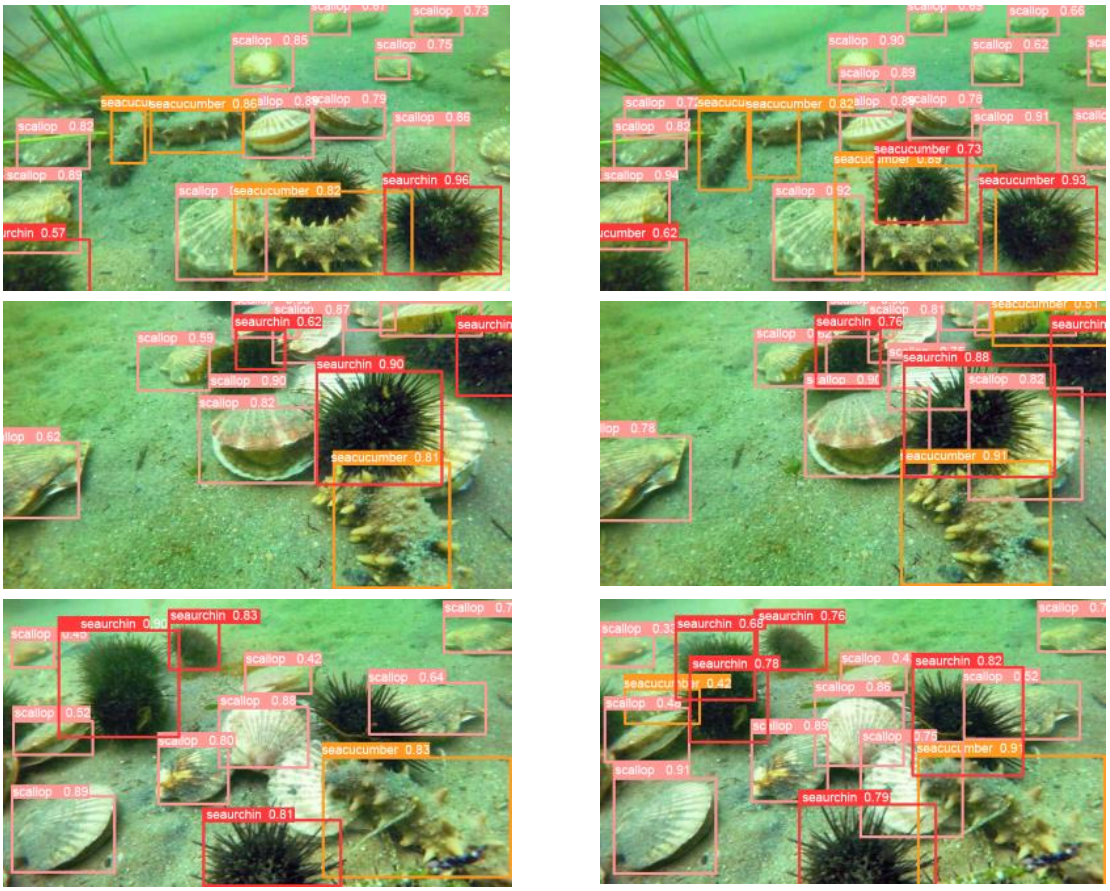

Figure 5: The detection results of the three sets of images, the left side is not added DN module, and the right side is added DN module

From the figure, we can see that from the overall point of view, the confidence level of the target detected using the model is relatively high, and the location information is more accurate. Partially overlapping, mutilated targets were not detected before the DN module was added, and the above problems were improved after we added the DN module. Therefore, without adding additional calculations, it can solve the problem of missing detection of ocean organism.

4. Conclusion

The parameters of the existing model of ocean organism target detection are relatively large, and it is not easy to deploy to the actual ocean organism breeding environment, and the SL-YOLO model proposed in this position can be similar to YOLOv5s in terms of detection accuracy, but the detection speed is much higher than that of YOLOv5s. Specifically: after replacing the feature extraction network of YOLOv5s with lightweight ShufflentV2, the mAP of the trained model decreased very little, but the amount of parameters decreased by 41.8%, indicating that the model is very suitable for the target detection model sought in this study, in order to balance the reduction in detection accuracy caused by the reduction of the number of parameters, after replacing the feature fusion part from PANet to the BiFPN module, the amount of model parameters has not been greatly improved. However, the accuracy of target detection can be improved by nearly 3 percentage points. Finally, in order to avoid the occurrence of missed detection due to target overlap, the use of DIOU_NMS instead of the original NMS algorithm, without changing the parameters of the model, can slightly improve the detection accuracy, and the situation of missed detection has also been improved.

ISSN: 2664-9640

References

[1] WHITE D J, SVELLINGEN C and ST R ACHAN N J C: Automatedmeasurement of species and length of fish by computer vision,Fisheries Research(2006), Vol.80 , p.203-210.

[2] FAN L Z,LIU Y:Automate fry counting using computer vision and multi－class least squares support vector machine, Aquaculture(2013),Vol.1 ,p. 91-98.

[3] Han F, Yao J, Zhu H, et al:Marine Organism Detection and Classification from Underwater Vision Based on the Deep CNN Method, Mathematical Problems in Engineering, 2020(2020), p.1-11.

[4] Dean Zhao, Xiaoyang Liu, et al: Underwater river crab identification method based on machine vision, Transactions of the Chinese Society for Agricultural Machinery(China 2019), Vol.50,p.151-158.

[5] Weicong Yu, Xianjiu Guo, et al: A marine treasure detection method based on lightweight deep learning Mobilenet-SSD network model, Journal of Dalian Ocean University(China 2021), Vol.36,p.340-346.

[6] Ma N, Zhang X, Zheng H T, et al: Shufflenet v2: Practical guidelines for efficient cnn architecture design, Proceedings of the European conference on computer vision (Munich, Germany, September 8-14, 2018). Vol.1,p.116.

[7] Wang K, Liew J H, Zou Y, et al: Panet: Few-shot image semantic segmentation with prototype alignment, Proceedings of the IEEE/CVF International Conference on Computer Vision(Long Beach, USA, June 15-20, 2019). Vol.1,p.9197.

[8] Tan M, Pang R, Le Q V. Efficientdet: Scalable and efficient object detection, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(Seattle, USA, Jun 14-19, 2020). Vol.1,p. 10781.

[9] Zheng Z, Wang P, Liu W, et al: Distance-IoU loss: Faster and better learning for bounding box regression, Proceedings of the AAAI Conference on Artificial Intelligence(New York, USA, February 7-12, 2020). Vol. 34,p.12993.

[10] Redmon J, Divvala S, Girshick R, et al: You only look once: Unified, real-time object detection, Proceedings of the IEEE conference on computer vision and pattern recognition(Harbin, China, August 7-10,2016). Vol.1,p. 779.

[11] Redmon J, Farhadi A: YOLO9000: better, faster, stronger, Proceedings of the IEEE conference on computer vision and pattern recognition(Washington, USA, September 6-9, 2017). Vol.1,p.7263.

[12] Redmon J, Farhadi A: Yolov3: An incremental improvement, arXiv preprint arXiv:1804.02767, 2018.

[13] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[14] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector, European conference on computer vision(Amsterdam, The Netherlands, October 8-16, 2016). Vol.1,p.21.

[15] Girshick R, Donahue J, Darrell T, et al: Rich feature hierarchies for accurate object detection and semantic segmentation, Proceedings of the IEEE conference on computer vision and pattern recognition(Tianjin, China, August 3-6, 2014).Vol.1,p. 580.

[16] Girshick R: Fast r-cnn, Proceedings of the IEEE international conference on computer vision(Beijing, China, August 2-5, 2015).Vol.1,p.1440.

[17] Ren S, He K, Girshick R, et al: Faster r-cnn: Towards real-time object detection with region proposal networks, Advances in neural information processing systems(2015),Vol.1, p.28.

[18] Dai J, Li Y, He K, et al. R-fcn: Object detection via region-based fully convolutional networks. Advances in neural information processing systems(2016),Val.1,p.29.

[19] Lin T Y, Dollár P, Girshick R, et al: Feature pyramid networks for object detection, Proceedings of the IEEE conference on computer vision and pattern recognition(Paris, France, May 21-25, 2017). Val.1,p.2117.

ISSN: 2664-9640

[20] Zhang X, Zhou X, Lin M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices, Proceedings of the IEEE conference on computer vision and pattern recognition(Chiago, USA, May 2-5, 2018). Val.1,p.6848.

[21] He K, Zhang X, Ren S, et al:Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition(Harbin, China, August 7-10,2016).Val.1, p.770.