ISSN:2790-1688 DOI: 10.56028/aetr.3.1.713

Underwater Target Detection Algorithm Based on Improved YOLOv5

Bin Ren[1,a], Jihe Feng[2], Yongdong Wei[2], Yuming Huang[2]

[1]International School of Microelectronics Dongguan University of Technology，Dongguan, China

[2]School of Computer Science and Technology Dongguan University of Technology，Dongguan, China

[a]renbin@dgut.edu.cn

Abstract. In order to solve the problem of low accuracy of traditional methods in underwater target detection and recognition, an underwater target detection algorithm based on YOLOv5, adding Convolutional Block Attention Module (CBAM) and Bidirectional Feature Pyramid Network (BIFPN) is proposed in this paper. In this method, CBAM is introduced into the output of YOLOv5 network model, and the attention module is used to fuse and enhance the features in channel dimension and spatial dimension, so as to enhance the features of target area features and improve the detection accuracy of small targets. At the same time, the simplified BIFPN module is used to replace the original enhanced feature extraction network in the Neck to improve the feature extraction ability of the network for different scales. The experimental results show that the mAP_0.5 is 84.2%, which is 0.9% higher than YOLOv5s model, meeting the needs of underwater target detection tasks.

Keywords: Target Detection; YOLOv5; CBAM; BIFPN

1. Introduction

With the development of science and technology, computer technology and its intelligent application have been widely used in various industries in national production and life. In the process of traditional fishery and aquaculture production, the commonly used fishing method mainly depends on the human eye to find the target. Therefore, for the fishing target located in the deep underwater and blind area of vision, the fishing efficiency of this method is not high. Thanks to the continuous development of marine technology and target detection technology in recent years, underwater target detection technology is also gradually popularized in aquaculture, fishing and other industries.

The most widely used method in traditional target detection technology is to recognize the object according to the characteristics of color, texture and edge. Among them, Hsiao[1] and others proposed a maximum probability local ranking method based on Sparse Representation-based Classification, which can well complete the task of fish recognition. Jing et al. [2] proposed a method of classifying fish based on texture and color features using a multi-class SVM. Joo et al. [3] extracted patterns of wild cichlids and used SVM and Random Forests for classification, achieving an accuracy of 72%. Although the above methods can complete the task well, they usually can only detect a single target, and require manual design program for feature extraction, resulting in large workload and low efficiency.

With the rise of artificial intelligence, more and more deep learning methods have been applied to underwater target detection. Goshorn et al. [4] used Haar classifier to detect butterfly fish. But their method is affected by the background and shooting angle in the image. Li et al. [5] improved the network structure of Fast R-CNN and proposed a light R-CNN suitable for underwater fish target detection, with an accuracy of 89.95%. Chuang et al. [6] proposed an underwater fish recognition framework based on completely unsupervised feature learning and error elastic classifier, which can better identify fish in different environments, with an average accuracy of 92.1%. Du et al. [7] proposed a fish recognition method based on multi-directional acoustic scattering data decision fusion based on SVM, and the recognition accuracy is more than 90%.

ISSN:2790-1688 DOI: 10.56028/aetr.3.1.713

The above methods based on deep learning technology have achieved good results in completing the target task, and have a high accuracy in detecting the situation of a single variety and specific scene. However, the effect is not very good for multi-species detection tasks. For solving the above problems and realize the underwater target fishing task with higher accuracy and better effect, this paper proposes an underwater target detection algorithm based on YOLOv5 [8][9], integrating the CBAM [10] Module and BIFPN [11] Module.

2. YOLOv5 Algorithm

This paper selects version 6.0 of YOLOv5s network model. Figure 1 shows the structure of YOLOv5 network. As can be seen from the Figure 1, YOLOv5 is mainly composed of Input, Backbone, Neck and Head. Among them, the Backbone network is a convolution neural network, which extracts feature of different sizes from the Input through multiple convolution and merging. The Neck network obtains more context information and reduces information loss by fusing these feature maps of different sizes. The Output part detects and classifies the target according to the new feature map generated by the Neck network. In YOLOv5 network architecture, CBS module is composed of convolution, normalization and SiLU activation function. The function of CSP (Cross-Stage Partial) module is to zoom out the model to improve the reasoning speed while maintaining the accuracy, which is distributed in the Backbone and Neck. In addition, the Spatial Pyramid Pooling Fast (SPPF) [12] module maximizes the pool through multiple 5*5 cores, and completes the fusion by connecting the features. The pooling layer operate through dimensionality reduction and represents features at a higher class of abstraction. It mainly compresses the input feature, which can reduce the computational complexity of the network and extract the main features. Concat module represents the concatenation operation of vectors.

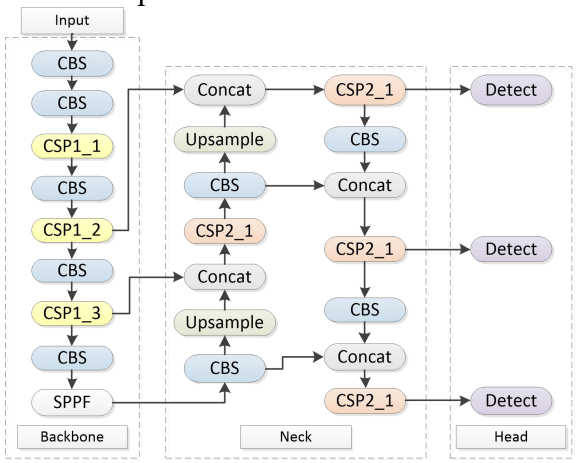

Figure 1. YOLOv5s network structure framework

3. Improved Algorithm Based on YOLOv5

3.1 CBAM Module

CBAM is an attention module for convolutional neural network. The module infers the attention map by two independent dimensions, and finally multiplies the attention map with the input feature for adaptive feature optimization. Compared with other attention modules, CBAM module has good applicability and low computing cost. Therefore, this paper uses CBAM module to further improve the performance of the algorithm.

In the Channel attention module, the input characteristic map is pooled to the maximum and average in parallel, so that attention can be better focused on the channel that has a greater impact on the final target, and then the compressed characteristic map is calculated on different scales of a shared full connection layer. Finally, the characteristic map enhanced by channel attention is output

ISSN:2790-1688 DOI: 10.56028/aetr.3.1.713

through Sigmoid activation function:

$$Mc(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) = \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{max}))) \quad (1)$$

In formula (1): σ is the Sigmoid activation function; MLP is a shared full connection layer; W_0 and W_1 are the input shared weight; F^c_{avg} and F^c_{max} represents the feature map generated in space by AvgPool and Maxpool respectively; C is the number of channels.

The Spatial attention module is responsible for paying attention to the position information of the meaningful part in the input feature. After Maxpool and AvgPool, the feature map F can better focus on highlighting the spatial features of the target and weaken the spatial features of other unrelated objects. The obtained two output features are spliced, convoluted by 7*7 convolution kernel, activated by σ, and finally output the feature map with spatial attention weight:

$$M_s(F) = \sigma\left(f^{7*7}([AvgPool(F); MaxPool(F)])\right) = \sigma\left(f^{7*7}\left([F^s_{avg}; F^s_{max}]\right)\right) \quad (2)$$

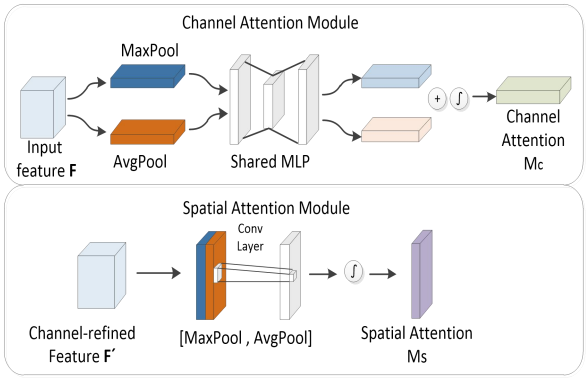

Figure 2. CBAM Module

3.2 BIFPN Module

BIFPN is a weighted bi-directional feature pyramid network proposed in EfficientDet algorithm [13]. It fuses feature maps of multiple scales. Different input features are not completely consistent with the importance of output features under different resolutions. The network is shown in Figure 3. In addition to forward propagation, the network reconstructs top-down and bottom-up two-way channels to fuse feature information from different scales of the Backbone network. Unifying the feature resolution scale by upsampling and downsampling, and adding double horizontal connections between features of the same scale, BIFPN alleviates the problem of feature information loss caused by too many network levels.

In feature extraction, the resolution of different input features is also different, and their contribution to the output features is obviously unequal. Therefore, BIFPN network adds additional learning weight to each input feature, so that the network can continuously adjust the weight to determine the importance of each input feature to the output feature.

ISSN:2790-1688 DOI: 10.56028/aetr.3.1.713

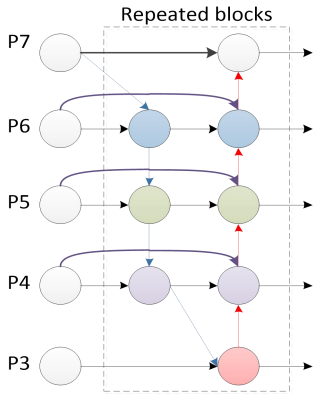

Figure 3. BIFPN module

3.3 Improved YOLOv5 Network

For the YOLOv5 Backbone, this paper retains the original network structure, and also extracts the features of the three feature layers of the Backbone network, and then transmits them to BIFPN network. However, since BIFPN network has five input feature layers, BIFPN network is simplified into three input feature layers to reduce the amount of calculation and adapt to YOLOv5 network. In addition, the improved algorithm in this paper introduces CBAM at the Output to strengthen the attention of the sent feature map, so as to further strengthen the network in the process of sending from the shallow layer to the deep layer. The learning of meaningful feature map, especially for small target features, can make the network better learn the feature information of the target, capture the recognition features of the target more accurately in the same test image, and achieve better recognition effect without increasing the training cost. The network structure of the improved algorithm is shown in Figure 4.

When the input size is (640,640,3), the three input feature layers of BIFPN module network are $P_3^{IN} = (80,80,128)$, $P_4^{IN} = (40,40,256)$, $P_5^{IN} = (20,20,256)$. Each feature fusion node of bifpn network will weight ω_i for each input feature respectively, and use the fast normalization formula to train these weights. The calculation formula of the output of each fusion node is shown in formula (3) to formula (6).

$$P_4^M = Conv\left(\frac{\omega_1 \cdot P_4^{IN} + \omega_2 \cdot Resize(P_5^{IN})}{\omega_1 + \omega_2 + \varepsilon}\right) \tag{3}$$

$$P_3^{OUT} = Conv\left(\frac{\omega_3 \cdot P_3^{IN} + \omega_4 \cdot Resize(P_4^{TD})}{\omega_3 + \omega_4 + \varepsilon}\right) \tag{4}$$

$$P_4^{OUT} = Conv\left(\frac{\omega_5 \cdot P_4^{IN} + \omega_6 \cdot P_4^{TD} + \omega_7 \cdot Resize(P_3^{TD})}{\omega_5 + \omega_6 + \omega_7 + \varepsilon}\right) \tag{5}$$

$$P_5^{OUT} = Conv\left(\frac{\omega_8 \cdot P_5^{IN} + \omega_9 \cdot Resize(P_4^{OUT})}{\omega_8 + \omega_9 + \varepsilon}\right) \tag{6}$$

In formulas (3) to (6): Conv represents convolution operation, Resize represents upsample or downsample operation in Input, $\omega_i \geq 0$ is a learnable weight, $\varepsilon = 0.0001$ is a small quantity to ensure numerical stability.

In the improved YOLOv5 structure, BIFPN network outputs three feature layers after enhanced feature extraction, and their shape sizes are respectively: $P_3^{OUT} = (80,80,128)$, $P_4^{OUT} = (40,40,256)$, $P_5^{OUT} = (20,20,256)$.

4. Experiment and Result Analysis

This experiment is evaluated on win10 system. The experiment uses the deep learning framework Pytorch1.10 and compiler Python3.6. The hardware environment is a computer with Intel i7-7820 and NVIDIA Titan XP.

4.1 Underwater Target Dataset

This paper adopts URPC2020 dataset, which contains 4 types of common benthos, with a total of 5543 pictures, but there are some errors in the position of the label and target frame of the dataset,

ISSN:2790-1688 DOI: 10.56028/aetr.3.1.713
so the data labels need to be repaired. After processing, 5042 valid pictures were obtained. The ratio of randomly selected training set to verification set is 8:2, including 4321 graphs in training set and 1081 graphs in verification set. The detection targets in the image are marked by Labelimg software, and the location and category information of all targets are recorded in TXT format. The training batch is set to 32 and the IOU[14] threshold is set to 0.5. All models train 150 epochs according to this parameter.

4.2 Evaluating Indicator

For evaluating the capability of the model, the precision rate (P), recall rate (R) and average precision (AP) are taken as the evaluation indexes, which are specifically expressed as follows:

$$P = TP/(TP + FP) \tag{7}$$

$$R = TP/(TP + FN) \tag{8}$$

$$AP = \int_0^1 P(R)dR \tag{9}$$

$$mAP = (\sum_{j=0}^{n} AP(j))/n \tag{10}$$

In formulas (7) to (10): TP is a positive example of successful prediction; FP is the negative case misjudged as positive by the model; FN is a positive case that is incorrectly predicted as a negative case by the model; AP (j) is the average accuracy of type j defects; j is the number of data set categories; mAP is the mean average precision.

4.3 Result Analysis

In order to verify the utility of the BIFPN module and CBAM module introduced in YOLOv5s, this paper uses YOLOv5s as the comparison algorithm. In addition, experiments are carried out to verify the impact of only introducing BIFPN module and CBAM module on the algorithm. The evaluation indicators mainly use AP and mAP, where AP represents the recognition accuracy of various targets, and mAP represents the average recognition accuracy of all targets. IOU=0.5 is a common standard for testing the performance of target recognition algorithms.

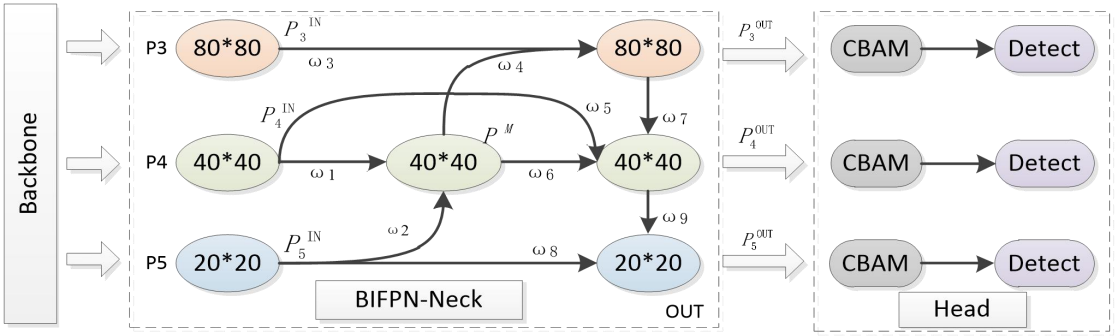

Figure 4. Improve YOLOv5 Network Structure

The experimental results are listed in Table 1.Compared with YOLOv5s, the mAP value obtained by YOLOv5s +CBAM that introduces the CBAM module in this paper is 83.9%, which is 0.6% higher than that of YOLOv5s, and in terms of AP value index, the holothurian is raise 0.7%. The scallop is raise 0.8%, and the starfish is raise 0.8%, which proves that the convolutional attention mechanism used in this paper can better get the target features, so the improvement is effective; the YOLOv5s+BIFPN which adds BIFPN module based on YOLOv5s, and its overall mAP value reached 84%, which is 0.7% higher than that of YOLOv5s, and in terms of AP value indicators, holothurian has increased by 1.9%, echinus has increased by 0.1%, scallop has increased by 0.6%, and starfish has increased by 0.4%, which proves the role of the BIFPN module. At the same time, the introduction of BIFPN module and CBAM module further improved the overall mAP value, and finally reached 84.2%. In terms of AP value indicators, holothurian increased by 1.9%, scallop increased by 0.3%, echinus increased by 0.4%, starfish increased by 0.6%, which proves that adding BIFPN module and CBAM module at the same time can bring better detection effect.

ISSN:2790-1688 DOI: 10.56028/aetr.3.1.713

5. Conclusion

In order to solve the problem of underwater target detection, this paper uses the more advanced YOLOv5 network. For improving the recognition effect of the algorithm, this paper uses a simplified BIFPN network to improve the recognition accuracy; In addition, CBAM module is added to effectively increase the extraction of important features and further strengthen the utilization of feature information by the network, which improved the effect of target recognition. The results show that the improved algorithm proposed is ideal for improving the accuracy. However, due to the limitations of the real marine environment image and the color of some creatures is similar to the environment, the recognition accuracy of some camouflage creatures needs to be further improved. In the future, we will explore new ways to further improve the recognition rate of targets similar to the ambient color.

Acknowledgment

This research has been supported by a number of project funds, including the Guangdong Provincial Basic and Applied Basic Research Fund Regional Joint Fundt (No. 2020B1515120095); Key Project of Dongguan Science and Technology of Social Development Program (No. 20211800904692, No.20211800904492); Dongguan Sci-tech Commissoner Program(No.2021180 0500012); Guangdong Provincial Enterprise Key Laboratory (No. 2020B121202001).

References

[1] Hsiao Y H , Chen C C , Lin S I , et al. Real-world underwater fish recognition and identification, using sparse representation[J]. Ecological Informatics, 2014, 23:13-21.

[2] Jing H , Li D , Duan Q , et al. Fish species classification by color, texture and multi-class support vector machine using computer vision[J]. Computers & Electronics in Agriculture, 2012, 88(none):133-140.

[3] Joo D , Kwan Y S , Song J , et al. Identification of Cichlid Fishes from Lake Malawi Using Computer Vision[J]. Plos One, 2013, 8.

[4] Goshorn D , Cho J , Kastner R , et al. Field Programmable Gate Array Implementation of Parts-Based Object Detection for Real Time Video Applications[C]// International Conference on Field Programmable Logic & Applications. IEEE Computer Society, 2010.

[5] Li X, Tang Y, Gao T. [IEEE OCEANS 2017- Aberdeen - Aberdeen, United Kingdom (2017.6.19-2017.6.22)] OCEANS 2017 - Aberdeen - Deep but lightweight neural networks for fish detection[J]. 2017:1-5.

[6] Chuang M C , Hwang J N , Williams K . A Feature Learning and Object Recognition Framework for Underwater Fish Images[J]. IEEE Transactions on Image Processing, 2016:1-1.

[7] Du W, Li H, Wei Y, Xu C.Decision fusion fish recognition method based on SVM [J]. Journal of Harbin Engineering University, 2015, 36 (05): 623-627.

[8] Redmon J, Farhadi A. (2018) YOLOv3: An incremental improvement. J. arXiv: 1804. 02767,2018.

[9] Redmon J , Divvala S , Girshick R , et al. You Only Look Once: Unified, Real-Time Object Detection[J]. IEEE, 2016.

[10] Woo S, Park J, Lee J Y, et al.CBAM: Convolutional Block Attention Module[C]// European Conference on Computer Vision. Springer, Cham, 2018.

[11] Lin T Y , Dollar P , Girshick R , et al. Feature Pyramid Networks for Object Detection[J]. IEEE Computer Society, 2017.

[12] He K , Zhang X , Ren S , et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[J].IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 37(9):1904-16.

[13] Tan Mingxing, Pang Ruoming and Le Quoc V. EfficientDet: Scalable and Efficient Object Detection[J].arXiv e-prints arXiv:1911.09070, 2019.

ISSN:2790-1688 DOI: 10.56028/aetr.3.1.713
[14] H Rezatofighi, Tsoi N , JY Gwak, et al. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019.

Table 1. Algorithm recognition accuracy comparison

Algorithm	AP/%(IOU=0.5)				mAP
	holothurian	echinus	scallop	starfish	
YOLOv5s	0.706	0.917	0.828	0.88	0.833
YOLOv5s+CBAM	0.713	0.917	0.836	0.888	0.839
YOLOv5s+BIFPN	0.725	0.918	0.834	0.884	0.84
YOLOv5s+CBAM+BIFPN	0.725	0.92	0.836	0.886	0.842