Research on an improved fish recognition algorithm based on YOLOX

Dongcai Liu[1], *Xianhui* Wen[1], and *Youling* Zhou[2,*]

[1]Hainan University, College of Information and Communication Engineering, China
[2]College of Information and Communication Engineering, Hainan University, Haikou, China

Abstract. The key to the development of underwater resources is to detect underwater targets quickly and accurately in real time. However, due to the influence of light, the underwater image is easy to be distorted and the contrast is low and so on, which greatly affects the performance of the detection algorithm, In order to improve the detection accuracy of underwater targets, After a detailed analysis of the underwater detection target features, The attention mechanism ECA module was added to the YOLOX model, Real-ESRGAN was used to treat multiple target and fuzzy images in detection images, the accuracy improved about 10 percent, A high-precision target detection algorithm suitable for underwater fish was developed, The ideal detection result was achieved.

Keywords: Deep learning, Underwater target detection, YOLOX, Attention mechanism ECA, ESRGAN.

1 Introduction

As the economy grows, Land resources are becoming increasingly limited, People turned their eyes under the water, But due to the complex underwater environment, be full of peril, Underwater robotics is becoming increasingly on the agenda, Traditional target detection method has insufficient identification accuracy, Deep learning technology has developed rapidly in recent years, Target detection based on deep learning is also one of the hot research topics, YOLOX is a relatively classical model in the single-stage target detection field, Among them, YOLOX is currently the most mainstream model in the field of target detection, Its analysis summarizes the key technologies in recent years, Integrated with other advanced detection techniques, And incorporate Anchor-Free into the model, Changing the traditional YOLO series based on anchor frame ideas.

Object detection is one of the most basic problems in computer vision. The region of interest in the image are identified and positioned, and it is widely used in autonomous driving, pedestrian detection, traffic object recognition, and license plate recognition.AI applications permeate all aspects of production and life, reducing the burden and facilitating people's lives. The current object detection algorithms have relatively high speed and

* Corresponding author: zhouyl@hainanu.edu.cn

accuracy, but in some fields, such as underwater targets and UAV aerial photography target identification, there are still insufficient detection accuracy.

Underwater environment is complex, collecting data difficult, image easy distortion, underwater target detection face much difficult, underwater robot technology development is increasingly mature, the underwater distance mainly rely on water acoustic image analysis, but water acoustic image analysis visualization ability is poor, only suitable for long distance large target positioning and tracking, underwater close target detection mainly depends on light visual image, image resolution, underwater close distance image detection has significant advantages. The development of underwater resources is of great significance for social and economic development and Marine defense construction. The excellent progress in YOLOX in the field of target detection in the past two years and YOLO (decoupling, data expansion, label allocation, Anchor-free mechanism, etc.), performance greatly improved, YOLOX set high performance and high speed. It has an important research significance to apply today's excellent object detection algorithm in underwater objective recognition. In this paper, firstly we use datasets (blue ring octopus, Sebastes and box jellyfish) make an experiment on YOLOX and find the problem that it have low accuracy .According to the experimental results, the ECA [18] module was embedded in the model, combined with the latest Real-ESRGAN[12], comparing the original YOLOX model, effectively improving the accuracy of underwater target detection and giving full play to the practicability of GAN network in the field of target detection.

2 Related work

Target detection first originated in face detection[1] in 1991.With the development of deep science technology and the improvement of computer hardware level, more and more target detection algorithms have emerged. The target detection algorithm is roughly divided into two stages: one is around 2000, mostly based on sliding window and manual extraction features, with large calculation volume and limited scope of application. The second phase, from 2014 to the present, began with RNN [2].It uses deep learning techniques to automatically extract features in images, then classification and prediction. Later, the target detection algorithms such as Fast R-CNN [3], Faster R-CNN [4], and SPP-Net [5] YOLO have further appeared. Compared with the traditional algorithms, the deep learning-based target detection algorithm has faster speed, high accuracy and strong generalization ability.

2.1 One-stage detection algorithm

One-stage's pioneering work was Redmon[6] et al that in 2016, YOLO (You Only Look Once) gave all the classification and regression tasks to the neural network, eliminating the selection of candidate boxes, faster, and truly realizing end-to-end target detection.In 2017, Redmon[7] et al. proposed YOLOv2, applying BN operations to each convolution layer, discarding Dropout, and accelerating the convergence of the model.Anchor boxes were used to predict the boundary boxes.The backbone network was replaced with Darknet-19 for the full connected layer in YOLO.In 2018 Redmon[8] et al proposed that YOLOv3 adopted the idea of feature fusion, with reference FPN using three different scales for target detection.Darknet-53 was used as the backbone network using a multi-label classification instead of Soft-Max.Binary cross-entropy was trained as a loss function to achieve the prediction of multiple categories in one box.In 2020, Bochkovskiy[9] et al. proposed that YOLOv4 used CSPDaeknet53 as the backbone network, introduced the Mosaic data augmentation method, and adopted the PANet network instead of FPN, to improve the detection accuracy of small targets.In 2021 Zheng Ge[10] et al proposed that the YOLO X backbone network adopted YOLO v3's Darknet53, YOLO X is equipped with some state-of-the-art detection techniques,

decoupling head, anchor-of-the-art label allocation strategies, better trade-off between speed and accuracy of all model sizes and all other peers. Note that increasing the YOLOv3 architecture to 47.3% of COCO is 3.0% higher than current best practices,

2.2 Image super-resolution based on deep learning

The SRCNN model proposed by Dong [12] et al. is the pioneering application of deep learning to image super-resolution. Through the convolution neural network, the image block extraction layer, nonlinear neural mapping and image reconstruction are completed.

Kim [13] VDSR model, the first residual structure for super-resolution reconstruction. The model network depth reaches 20 layers, and the deeper the network structure has a greater receptive field.

2.3 Attention mechanism

In 2018 Hu et al [16] proposed that SENet adaptively recalibrates channel-level feature responses by establishing interdependencies between channels, and SE blocks produce significant performance improvements on existing deep network architectures with minimal additional computational cost.

In 2020, wang [17] et al. proposed the ECA module, which does not require a dimension-reduced local cross-channel interaction strategy, as effectively implemented via one-dimensional convolution. The module involves only a small number of parameters, while bringing in a significant performance gain. Through SENet channel attention studies, avoiding dimensional reduction is important for learning channel attention, and appropriate cross-channel interactions can maintain performance while reducing model complexity. In addition, an adaptive selection of the kernel size of one-dimensional convolution was developed to determine the local cross-channel interaction coverage.

3 Methed and materials

At present, several important parameters used in the evaluation of performance of object detection model are mAP (mean Average Precision), average accuracy mean and AP (Average Precision). It is the main evaluation index of target detection algorithm. Object detection model usually describes the model with speed and accuracy (mAP) indicators. The higher the mAP, the better the detection effect of the detection model on a specific data set.

$$\mathrm{IoU}(b_{\mathrm{pred}}, b_{gt}) = \frac{\mathrm{Area}\ (b_{\mathrm{pred}} \cap b_{\mathrm{gt}})}{\mathrm{Area}\ (b_{\mathrm{pred}} \cup b_{\mathrm{gt}})} \tag{1}$$

The intersection ratio threshold (IoU Threshold) is a predefined constant expressed as Ω when $\mathrm{IoU}(b_{\mathrm{pred}}, b_{gt}) > \Omega$ considered b_{pred} a positive sample (including an object) and otherwise a negative sample (background).The accuracy and recall of target detection can be calculated based on the intersection ratio and intersection ratio threshold:

$$\mathrm{Pre}cision = \frac{TP}{(TP + FP)} \tag{2}$$

$$\mathrm{Re}call = \frac{TP}{(TP + FN)} \tag{3}$$

Where TP, FP, FN are the true positive rate, false positive rate, and false negative rate, respectively, indicating the number of positive samples being predicted correctly, the number of negative samples being predicted as positive samples and the number of background being false as positive samples. The accuracy and recall were calculated, and the threshold Ω =0.5.

Average accuracy rate (Average Precision, AP) is also a commonly used index, with the formula of:

$$\mathrm{AP} = \int_0^1 \mathrm{Presion}(t)dt \tag{4}$$

Where Presion (t) represents the exact rate at the threshold Ω =t. For multi-category detection tasks, because different categories may exist in the image to be detected, Mean Average Precision (mAP) is usually used as the evaluation index, and the formula is as follows:

$$\mathrm{mAP} = \frac{\sum_{n=1}^{N} AP_n}{N} \tag{5}$$

N is the number of object categories, and AP_n represents the accuracy of the algorithm for evaluating the objects in the n category.

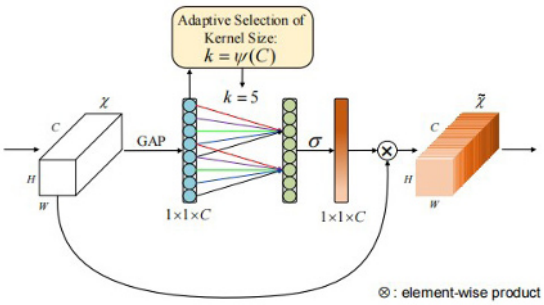

Figure 1 ECA model.

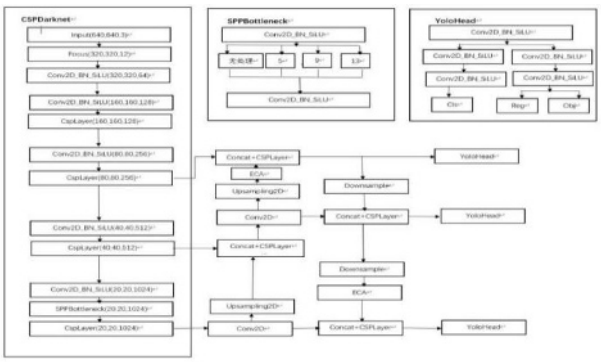

Fig. 2. The YOLOX model architecture added to the ECA module.

As shown in Figure 1, given the aggregate features obtained by global average pooling (GAP), the ECA generates the channel weights by performing a fast one-dimensional convolution of size k, where k is adaptively determined by the mapping of the channel dimension C. After performing channel-level global mean pooling without dimension reduction, the ECA captures local cross-channel interactions by considering each channel and its k neighbors. ECA can be efficiently achieved by fast one-dimensional convolution of size k, where the kernel size of k represents the coverage of local cross-channel interactions, namely how many neighbors are involved in the attentional prediction of a channel. To avoid manual tuning of k by cross-validation, an adaptive method for determining k was developed, where the coverage of the interaction (i. e., the kernel size k) is proportional to the channel dimension.

As shown in Figure 2, the attention ECA module was added before the upper and down sampling features were fused.

4 Experients

4.1 Experimental result

This paper collected 150 underwater data sets with three categories: blue ring octopus, Sebastes and box jellyfish, with 50 pictures in each category. Pre-training weights were loaded at the beginning of the training, and tested on YOLO X.As shown in Figure 3, the trained loss curves.

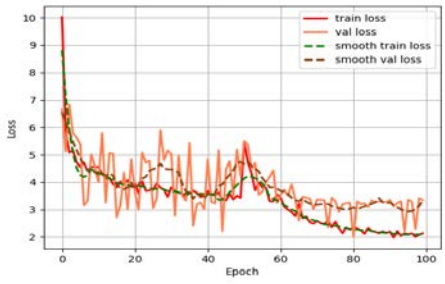

Fig. 3. YOLOX Loss curve.

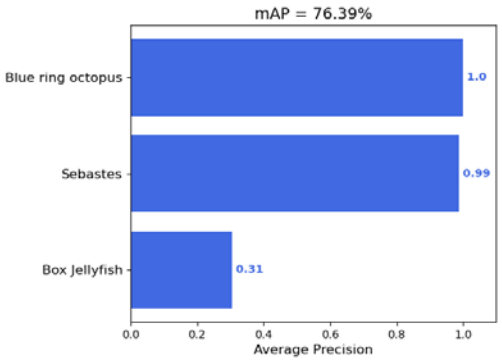

Fig. 4. Mean accuracy of the YOLOX assay.

As can be seen from Figure 3, the loss decline of YOLOX is relatively gentle, and the training time is longer.About 100 loss losses can be close to minimum.

4.2 Result analysis

As can be seen from Figure 4, the average accuracy of YOLO X is 76.39%. Due to the number of randomly assigned images when validation is too low, the accuracy of Box Jellyfish validation is too low, but when the training weights are loaded at the end of training for validation, the accuracy is much higher than this, at about 90%.

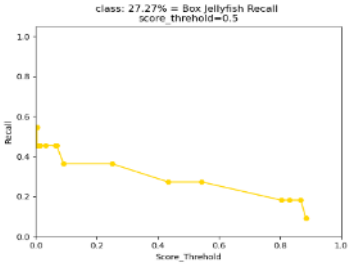

Fig. 5. Recall rate of box-jellyfish–YOLOX.

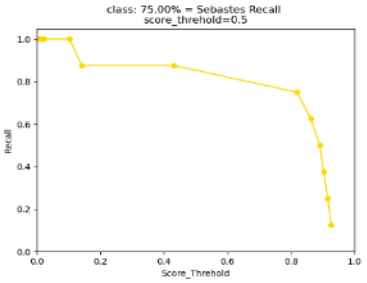

Fig. 6. The recall rate of sebastes-YOLOX.

Recall is for the original sample, which indicates how many of the positive cases in the sample are predicted correctly. The figure shows the recall rate corresponding to the threshold confidence is 0.5 when predicted. The recall curve of the sebastes is still normal. Because the training set and test set are divided according to the proportion of the original data set, the total data set is small, and there are also uneven types of samples used for the test.

Based on the above experimental analysis, YOLOX still has some room to improve in the underwater fish detection. In view of the above problems, this paper uses ESRGAN to enhance the prediction data, and inserts the ECA attention module in the YOLOX model to further improve the accuracy of the final detection results and make it more suitable for the detection of underwater targets.

4.3 Yolox and reall-esrgan

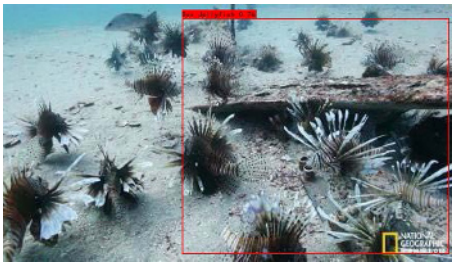

Fig. 7. Detection results of the Sebastes on the YOLOX model.

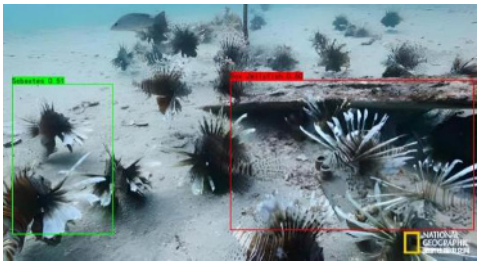

Fig. 8. Detection results of the YOLOX model after super-resolution processing with ESRGAN.

The detection results of FIGS. 7 and 8 show that only the wrong target can be detected in no image processing, and the target in the image is correctly detected immediately after the super-resolution of the image. Moreover, the false detection rate can be effectively reduced. Since the less multi-target data when training the model and the limited data set, the model is not good at detecting the multi-target in the picture. However, this does not affect the improvement of ESRGAN detection effect on model detection.

Fig. 9. Detection results of the Sebastes on the YOLOX model.

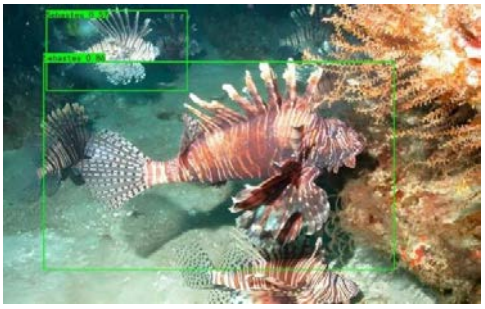

Fig. 10 Detection results of YOLOX after super-resolution processing with ESRGAN.

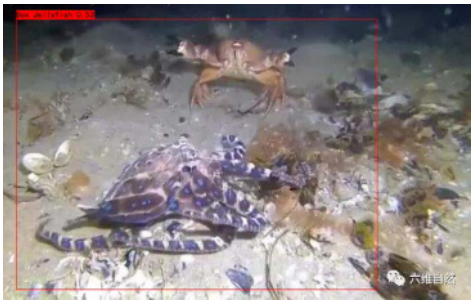

Fig. 11. Results of the Blue Ring Octopus on the YOLOX model.

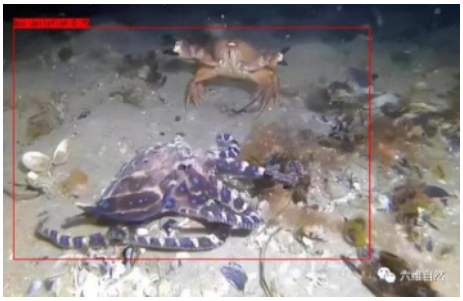

Fig. 12. YOLOX detection results after super- resolution processing with ESRGAN.

The results of FIGS. 11 and 12 show that ESRGAN also improves the detection accuracy of the detection model.

4.4 Yolox and reall-esrgan and eca

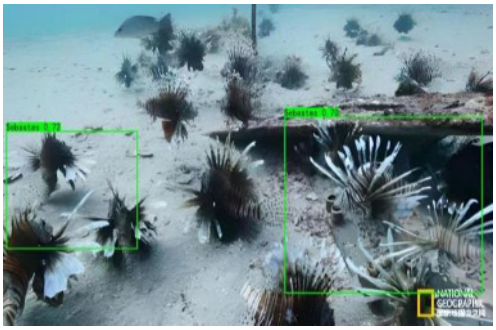

Fig. 13. YOLOX The addition of the ECA module and combining the detection results of ESRGAN.

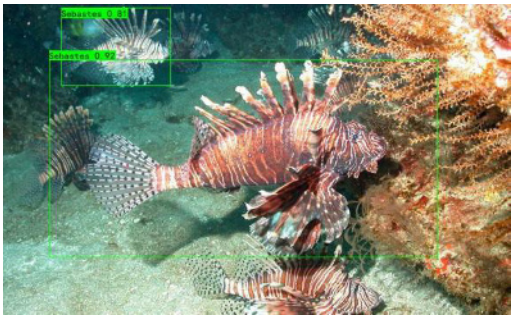

Fig. 14. YOLOX the addition of the ECA module and combining the detection results of ESRGAN.

From FIGS. 13, 14, the overall accuracy of the attention module in the model is greatly improved, and from FIGS. 8 and 13 that the original wrong detection results were corrected after the ECA is added, and the detection accuracy is improved, mainly due to the location of the addition. After the analysis of multiple experimental results, the attention module is significantly effective before the feature fusion at the output end of the backbone network.

5 Conclusion

This paper applies the current mainstream target detection model YOLOX to underwater target detection. After experimental analysis, it can be seen that YOLOX model still has great

room for improvement in underwater fish detection. For YOLOX in underwater target detection, this paper adds ECA module to YOLOX model, and ESRGAN super resolution processing for the detection image can improve the accuracy of model detection and correct the detection results., in the future in practical application more on several models, combined with cost and speed to find the most appropriate model can perfect their actual task.GAN network developed rapidly in recent years, this article using GAN image super resolution, expand the target in the image, multiple target detection and detection accuracy are improved, but after super resolution processing increases the computing cost, super resolution processing also need to see, generally just for multi-target or small target, if the detection accuracy super resolution, obviously outweigh the loss, combined with experimental analysis can be seen in low detection accuracy and more target can use GAN technology for super resolution processing.in the future can explore change the backbone of the model and enhance the model feature extraction ability to improve detection accuracy.

This paper was supported by National Natural Science Foundation of China (grant number 61340027)

References

1. Turk M A, Pentland A P. Recognition in face space[C]//Intelligent Robots and Computer Vision IX: Algorithms and Techniques. International Sociey for Optics and Photonics, 1991, 1381: 43-54.
2. Gkioxari G, Hariharan B, Girshick R , et al. R-CNNs for Pose Estimation and Action Detection[J]. Computer ence, 2014.
3. Girshick R. Fast R-CNN[J]. arXiv e-prints, 2015.
4. Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. 2017.
5. Purkait P, Zhao C, Zach C. SPP-Net: Deep Absolute Pose Regression with Synthetic Views[C]// British Machine Vision Conference(BMVC 2018). 2017.
6. Redmon J, Divvala S , Girshick R , et al. You Only Look Once: Unified, Real-Time Object Detection[J]. IEEE, 2016.
7. Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2017:6517-6525.
8. Redmon J, Farhadi A . YOLOv3: An Incremental Improvement[J]. arXiv e-prints, 2018.
9. Bochkovskiy A, Wang C Y , Liao H . YOLOv4: Optimal Speed and Accuracy of Object Detection[J]. 2020.
10. Ge Z, Liu S, Wang F, et al. Yolox: Exceeding yolo series in 2021[J]. arXiv preprint arXiv:2107.08430, 2021.
11. Wang X, Xie L , Dong C , et al. Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data[J]. 2021.
12. Dong,Image Super-Resolution Using Deep Convolutional Networks. 2016.
13. J Kim, Shin M J, D Kim, et al. Performance Comparison of SRCNN, VDSR, and SRDenseNet Deep Learning Models in Embedded Autonomous Driving Platforms[C]// 2021 International Conference on Information Networking (ICOIN). 2021.
14. Radford A, Metz L,Chintala S.Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks[J]. Computer ence, 2015.

15. Kim J, Lee J K, Lee K M. Accurate Image Super-Resolution Using Very Deep Convolutional Networks[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2016.
16. Jie H, Li S , Gang S . Squeeze-and-Excitation Networks [J]. IEEE, 2018.
17. Wang Q, Wu B, Zhu P, et al. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.
18. Liu Y, Wang Y, Wang S, et al. CBNet: A Novel Composite Backbone Network Architecture for Object Detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7):11653-11660.