

Research on Detection Method of Daily Necessities based on YOLOv5s

Yicheng Wang^{1,2}, Guoliang Zhang^{1,2,*}

¹ School of Automation and Information Engineering, Sichuan University of Science and Engineering, Yibin 644000, China

² Artificial Intelligence Key Laboratory of Sichuan Province, Yibin 644000, China

* Corresponding author: Guoliang Zhang (Email: zhgl@sohu.com)

Abstract: A daily necessities detection algorithm based on improved YOLOv5 is proposed to address the issue of low detection accuracy in daily necessities. Firstly, replace the Euclidean distance of the original K-means clustering algorithm with a 1-IOU approach to detect anchor boxes that are more closely aligned with the target object. Secondly, in order to facilitate and quickly fuse multi-scale features in the network, BiFPN is used in the Neck layer to replace the original FPN+PAN structure. Then, in order to reduce the interference of the background environment on the algorithm, improve the performance and generalization ability of the model, the ECA-Net attention mechanism is introduced after BiFPN. Finally, the feasibility of the proposed method will be verified through detection experiments supported by the self-made dataset. The experimental structure shows that the improved method can improve the detection accuracy of daily necessities, with an accuracy increase of 3.30% compared to the original YOLOv5 algorithm.

Keywords: YOLOv5; Object Detection; K-means Clustering; ECA-Net; BiFPN.

1. Introduction

With the rapid development of object detection, it has been widely applied in industries [1], agriculture [2], medical and other fields. The emergence of object detection has alleviated the problems of low production efficiency, lack of labor, and safety risks in these fields. However, there are still problems such as missed detection and false detection when facing most quantities and different types of objects.

The algorithms for object detection are divided into two categories: two-stage object detection and single-stage object detection. The two-stage detection algorithm needs to find candidate regions in the feature map and perform classification and regression by sliding a window on the candidate regions. Girshick R [3] et al. proposed the R-CNN (Region Convolutional Neural Network) detection algorithm as a representative of two-stage detection algorithms, with an accuracy MAP of 66.9% on the PASCAL VOC dataset. Ren S et al. proposed the Faster R-CNN algorithm based on the VGG-16 network, further improving the R-CNN algorithm. This algorithm uses Region Proposal Network (RPN) instead of Selective Search (SS), and can filter candidate regions through RPN calculation, reducing the number of candidate regions from around 2000 to around 300, greatly improving detection speed and efficiency. Zhang Yi et al. [5] improved on Faster R-CNN by using multi-dimensional feature fusion to improve the detection requirements for small targets, and conducted experimental verification on the TT100K dataset. The detection accuracy reached 90.83%, but its computational complexity and real-time performance were poor. Pei Yingling et al. proposed an improved Faster R-CNN object detection algorithm, which introduces deformable convolution in the feature extraction network to construct deformable residual convolution blocks and uses Alpha-IoU as the target regression loss function to improve the accuracy of object detection. The single stage detection algorithm directly uses regression methods to detect and classify target objects. Compared to the two-stage algorithm, it has better

real-time performance and poor detection accuracy due to the lack of "one stage" steps. Redmon J et al. proposed the YOLOv3 object detection algorithm, which replaces the original Darknet-19 backbone feature extraction network with the DarkNet-53 backbone feature extraction network based on the YOLOv2 network. The algorithm uses multi-scale features to detect objects, and uses cross entropy functions to predict the categories of detected objects, effectively improving detection accuracy compared to previous versions. Bochkovskiy et al. [8] optimized the YOLOv3 algorithm and proposed the YOLOv4 algorithm by adding spatial pyramid pooling (SPP) and path aggregation network (PAN) to further improve the network's feature extraction ability. Finally, improvements were made to the input during training to achieve good performance on general GPUs. Niu Xinyu et al. [9] improved the YOLOv5s algorithm by using ShuffleNetv2 as the feature extraction network and adding a CA attention mechanism on top of the backbone network to more accurately detect the target area of interest. At the same time, GSConv and VOV-GSCSP modules were used in the Neck layer to reduce the complexity of the model. Yang Guoliang et al. [10] introduced the Kmeans++ clustering algorithm to improve the positioning accuracy of target objects for different scales of tomatoes, and added a GAM attention mechanism at the end of the backbone network to improve the detection accuracy of target objects. At the same time, BiFPN weighted bidirectional feature pyramid results were used to replace the original FPN+PAN structure to obtain richer features.

The above detection methods do not have efficient detection capabilities on the basis of real-time detection, and are not suitable for daily life detection tasks. Moreover, there is limited research on the detection of daily necessities. This article will propose a YOLOv5s detection algorithm based on daily necessities, replacing the original FPN+PAN structure with BiFPN bidirectional weighted pyramid structure and adding ECA-Net attention mechanism.

2. YOLOv5 Object Detection Algorithm

The YOLOv5 object detection algorithm only requires a glance to know the detected category. YOLOv5 has a total of four models, sorted by complexity: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5n. Due to the fact that the YOLOv5s

model has the lowest complexity among the four models and is relatively simple to implement, this article adopts the YOLOv5s network model for improvement. The YOLOv5s model is mainly composed of four parts, namely the input layer, feature extraction layer (Backbone), feature fusion layer (Neck), and input layer (Head). The network structure diagram is shown in Figure 1.

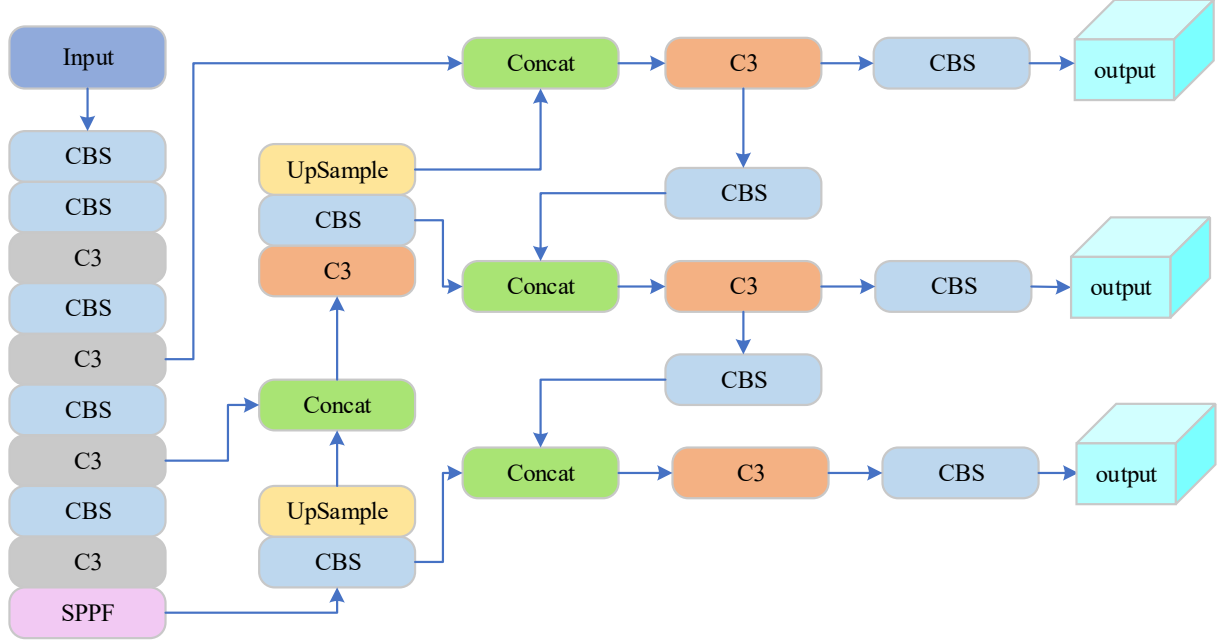


Figure 1. YOLOv5s Network Structure Diagram

(1) Backbone

The benchmark network Backbone uses the CSP Marknet53 structure, mainly composed of Convolutional Batch normalization Silu Block (CBS), bottleneck module (C3), and Spatial Pyramid Pooling Fast (SPPF). In the input stage of the network, a convolutional kernel size of 6 is used $d \times 6$. The convolutional structure with a step size and fill of 2 replaces the Focus structure before 5.0, further improving computational efficiency and reducing computational costs while maintaining the previous Focus structure for double downsampling. This convolutional structure is shown in Figure 2.



Figure 2. CBS Convolutional Structure

(2) Neck

The Neck layer is located between Backbone and Head, and adopts the structure of Feature Pyramid Networks (FPN) and Path Aggregation Network (PAN). The FPN structure uses a top-down approach to transmit the semantic feature information from the upper layer to the lower layer, allowing the model to obtain richer semantic information. The PAN structure enhances the model's acquisition of positional information by using a bottom-up approach. The combination of the two structures further improves the feature extraction ability of the network model. The FPN+PAN structure diagram is shown in Figure 3.

(3) Head

The Head layer consists of two parts: bounding box and non-maximum suppression (NMS). In early versions of YOLOv5, functions such as IOU Loss and GIOU Loss were

used. The YOLOv5 6.0 version used in this article uses CIOU_ as the loss function of the Bounding box, Loss can better solve the problem of overlap between the predicted box and the real box compared to the previous loss function used.

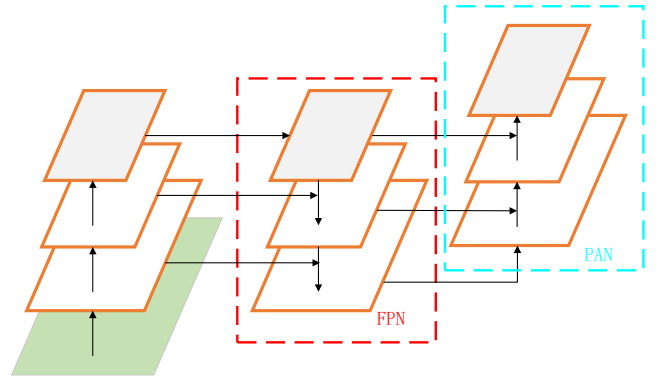


Figure 3. FPN+PAN Structure Diagram

In daily detection tasks, due to issues such as poor trained models or insufficient datasets, an object may correspond to multiple detection boxes. In this case, the NMS function selects the highest scoring anchor box through a scoring mechanism based on the detected anchor box, and then redundant other lower scoring anchor boxes to select the optimal detection box.

3. Improved YOLOv5 Algorithm

3.1. K-means Clustering Algorithm

K-means clustering algorithm, where K represents the number of K clusters, and means represents the average value of the clusters in the numerical value. This value is taken as the cluster center point, which is the cluster center. The

clustering diagram is shown in Figure 4. The implementation process of this algorithm is as follows:

- (1) Firstly, determine the number of K clusters based on the sample category.
- (2) The Euclidean formula (Euclidean distance) is generally used to calculate the distance from the sample point to the center of the cluster.
- (3) Reclassify cluster centers based on calculation results.
- (4) Repeat operations (2) and (3) until the cluster center does not change.

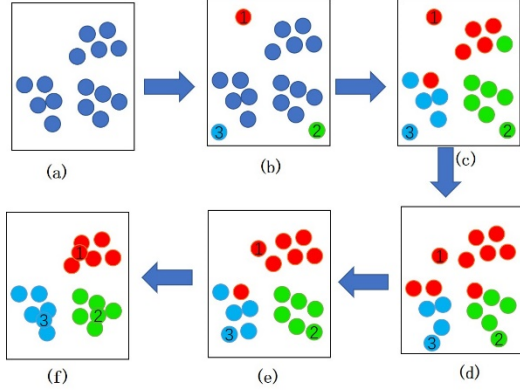


Figure 4. K-means clustering algorithm flowchart

The original GG-CNN network takes the original depth map of the target object as input, encodes the depth information through simple convolution and deconvolution, and decodes the captured information. The algorithm itself does not preprocess the input data, and the lightweight model structure limits the network's feature extraction ability and attention to capturing sensitive areas. To address the above issues, a series of improvements were made to the original GG-CNN model to enhance the algorithm's grasp detection ability from multiple perspectives. The improved GG-CNN is shown in Figure 2, with the red dashed lines in the figure indicating the improvement points.

3.1.1. Optimizing K-means Clustering Algorithm

In common engineering tasks, the YOLOv5s model sets 9 anchor values, namely 9 cluster centers, and allocates anchor sizes reasonably based on the scale of feature maps. The allocation principle is that large-scale feature maps correspond to smaller anchor values, mesoscale feature maps correspond to lower anchor values, and small-scale feature maps correspond to larger anchor values. Considering that the original K-means clustering algorithm used Euclidean formula to measure the distance between Anchor Box and Ground Truth would result in more errors, a 1-IoU method was used to replace the Euclidean distance in the original K-means clustering algorithm. The optimized anchor box size is shown in Table 1.

Table 1. Optimized anchor box size

Feature map size/pixels	Anchor box size/pixels
20×20	[298,316], [577,202], [585,359]
40×40	[230,69], [175,150], [433,120]
80×80	[48,17], [117,31], [70,70]

3.2. Improvement of Backbone Networks based on Attention Mechanism

This chapter takes into account the large size and complex background of the self-made dataset, the slow convergence of the model during the training process, and the susceptibility of the target to be detected in subsequent detection and grasping experiments to background interference, leading to

issues such as false detections and missed detections. Therefore, using the PANet structure idea to replace the original FPN+PAN structure with a weighted bidirectional pyramid structure (BiFPN) for multi-scale feature fusion, in order to obtain richer semantic information; Add the ECA-Net attention mechanism to the YOLOv5 network, strengthen the relationship between the model and channels, improve the performance and generalization ability of the model, and accelerate the training of the model.

In order to better improve the detection performance of targets at different scales, a more efficient bidirectional feature pyramid network (BiFPN) structure can be used to quickly fuse multi-scale features. Compared with the PANet structure, BiFPN removes intermediate nodes to simplify the network and reduce computational complexity, and adds skip connections between input and output points, in order to integrate more features while saving computational costs, the structural diagram is shown in Figure 5.

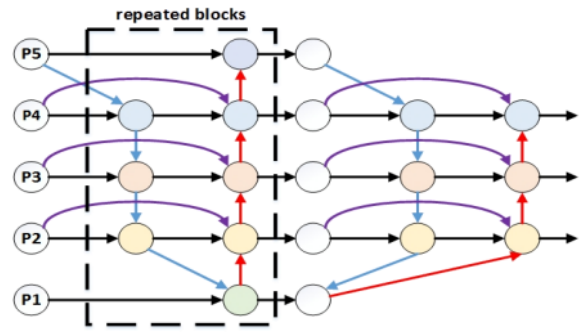


Figure 5. Structure of BiFPN [11]

The Efficient Channel Attention for Deep Convolutional Neural Networks (ECA-Net) is an improved version of the SENet channel attention mechanism. The SENet attention mechanism uses a fully connected layer (FC) structure, which can lead to an excessive number of parameters and affect training results. ECA-Net replaces the fully connected layer with a one-dimensional convolution with a convolutional kernel size of after the Average Pooling layer, achieving cross channel information fusion, reducing the number of parameters, and achieving good results. The network structure is shown in Figure 6.

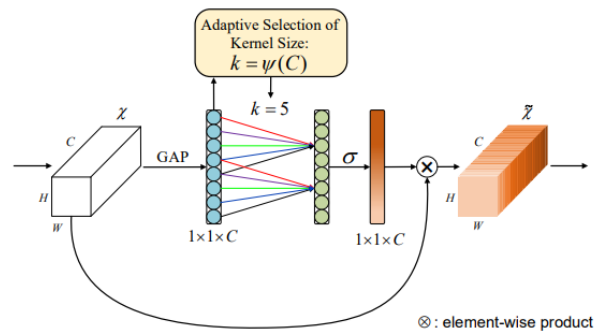
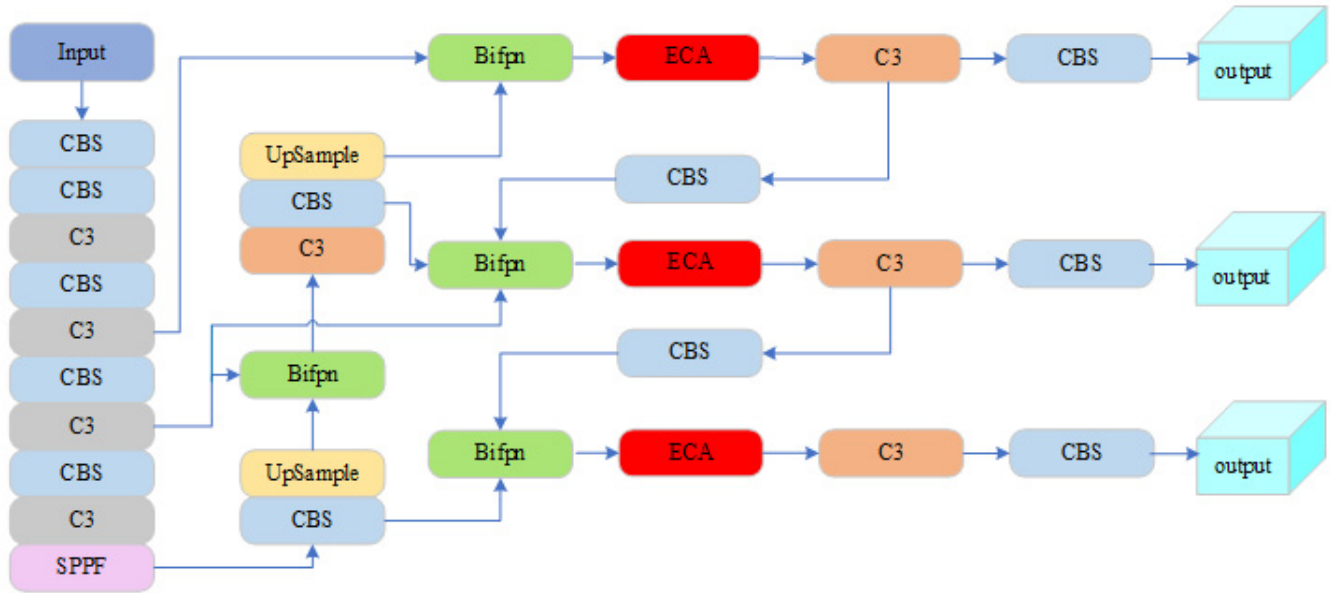


Figure 6. ECA Net [12] Structure Diagram

Due to the fact that most existing algorithms focus on improving complex attention, it is inevitable to increase the complexity of the model. In order to provide more detailed information in areas that need attention and achieve better performance while maintaining the original lightweight advantage, this chapter embeds the ECA Net attention mechanism into the YOLOv5s model. The addition method is shown in Figure 7, which can accelerate the convergence

speed of the model.



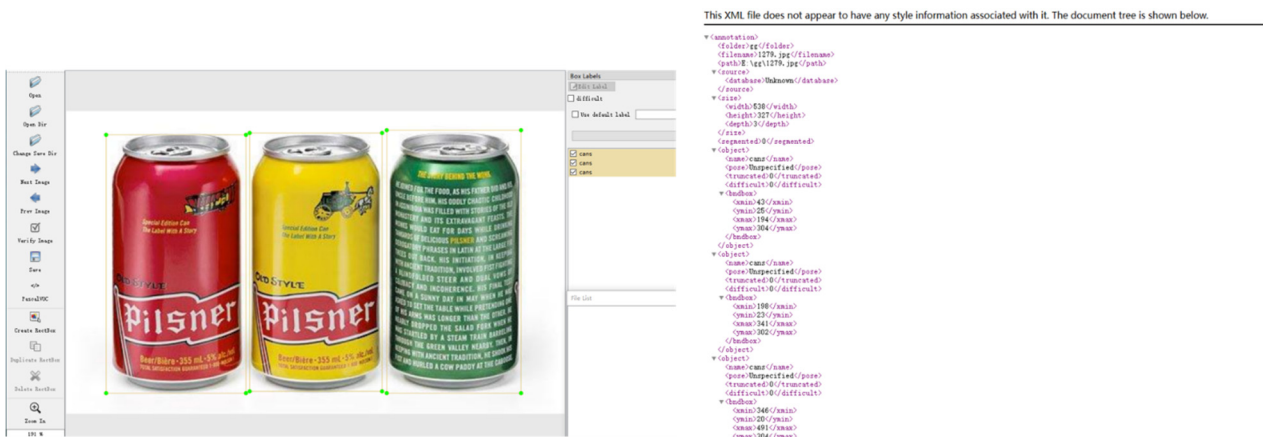
4. Experimentation

Considering that the dataset categories used in this article are limited or even absent from numerous publicly available target detection datasets, and the equipment conditions used in the experiment cannot train too large a dataset, a self-made target detection dataset was created in the YOLO format.

4.1. Self-made Dataset Process

The self-made target detection dataset image in this article consists of two parts: one is downloaded from Baidu, and the other is captured using the Intel Realsense D435i visual sensor and annotated using the commonly used annotation tool Labeling [13] for target detection. As shown in Figure 8

(a), taking the annotation can as an example, import the image that needs to be annotated into Labelimg. Click on the left taskbar to select the object in the box, and add a label name to it, and finally keep the annotated image in the specified position. After the annotation is completed, an XML format file as shown in Figure 8 (b) will be obtained, which will display information such as the label name and the width and height of the box. Through this annotation method, each image is annotated in a total of 5 categories: cans, remote control, shoes, cell phone, and bottle. There are a total of 5511 images in the dataset, which are 901, 837, 1163, 1018, and 1592 in order of number. There are a total of 8676 annotation boxes, which are 2031, 1417, 2403, 1417, and 1661 in order of number. The dataset is divided into training, validation, and testing sets in a ratio of 8:1:1.



4.2. Experimental Configuration

The operating system used in this chapter is Windows 10; Software configuration Anaconda, Pycharm; CUDA11.1, CUDNN8.0.5 to support GPU; The deep learning framework is Pytorch 1.7.1. The hardware configuration is shown in Table 2, and the hyperparameters used during training are shown in Table 3.

Table 2. Experimental Configuration Platform

Name	Model
CPU	RTX3060 12GB
GPU	i5-12600KF
Memory	32G
motherboard	ASUS H610M-K D4

Table 3. Partial hyperparameters

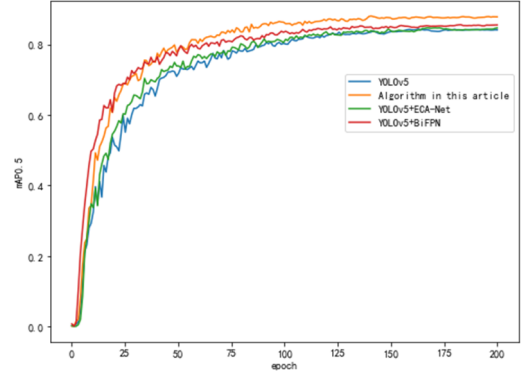
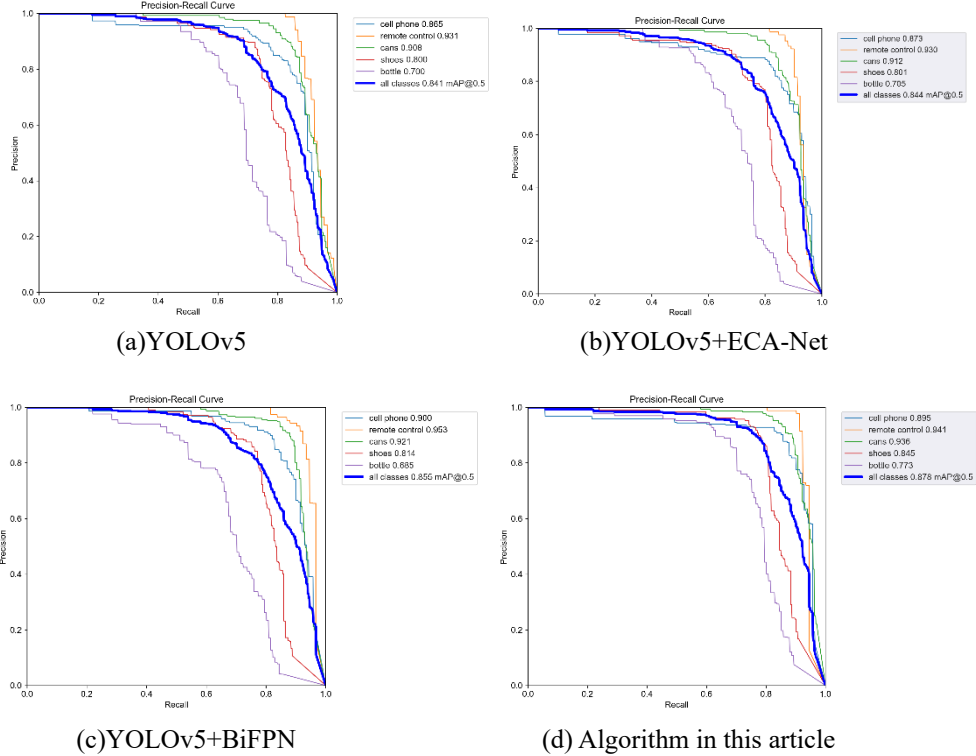
Parameter	Price
Input shap	640×640
Batch	2
Epoch	200
learning rate	0.01
momentum	0.937
weight decay	0.12

4.3. Training Process

Based on the above experimental configuration, the image is displayed at $640 \times$. The size of 640 is input into the network for training. In response to the proposed improvement points in Section 2, the self-made dataset is placed into the YOLOv5 algorithm before and after the improvement for training. The trained mAP accuracy diagram is shown in Figure 9.

In Figure 9, it can be seen that the accuracy of YOLOv5, YOLOv5+ECA-Net, YOLOv5+BiFPN, and the improved algorithm mAP in this article are 84.08%, 84.44%, 85.46%, and 87.78%, respectively. The mAP accuracy of YOLOv5 improved in this article has increased by 3.70% compared to the original YOLOv5, 3.34% compared to YOLOv5+ECA-

Net before improvement, and 2.32% compared to YOLOv5+BiFPN. In order to more intuitively see the AP rise and fall of each category, a P-R curve diagram is drawn based on the trained data, as shown in Figure 10.

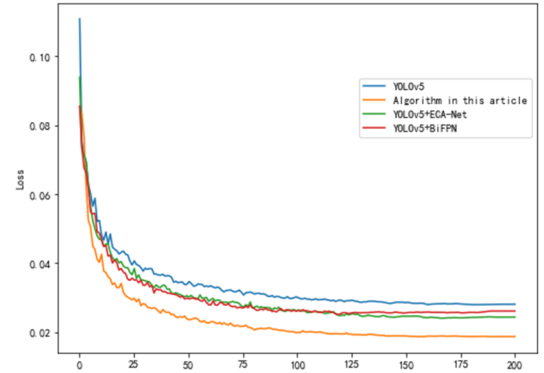
**Figure 9.** Comparison of mAP of YOLOv5s model before and after improvement**Figure 10.** P-R curves of various algorithms

From the loss graphs of various algorithms in Figure 11, it can be seen that the YOLO series of algorithms were trained with 200 Epochs, and their convergence performance significantly improved when only ECA-Net attention mechanism was added. The convergence performance was slightly better when only BiFPN was added than the original YOLOv5. After fusing the above two improved points together, the convergence performance improved and the loss value region remained stable after 125 Epochs.

4.4. Ablation Experiment

In order to verify the effectiveness of the improved YOLOv5 object detection algorithm proposed in this article, self-made datasets were placed into 6 categories of algorithms, including Faster RCNN, SSD, and YOLOv5s object detection

algorithms before and after improvement.

**Figure 11.** Comparison of YOLOv5 Loss Chart before and after Improvement

Training was conducted on the same experimental platform and detection comparisons were conducted. The comparison results are shown in Figure 12. Train and compare the self-made dataset into 6 types of algorithms, including Faster

RCNN, SSD, and YOLOv5s object detection algorithm before and after improvement. The comparison results are shown in Figure 12.

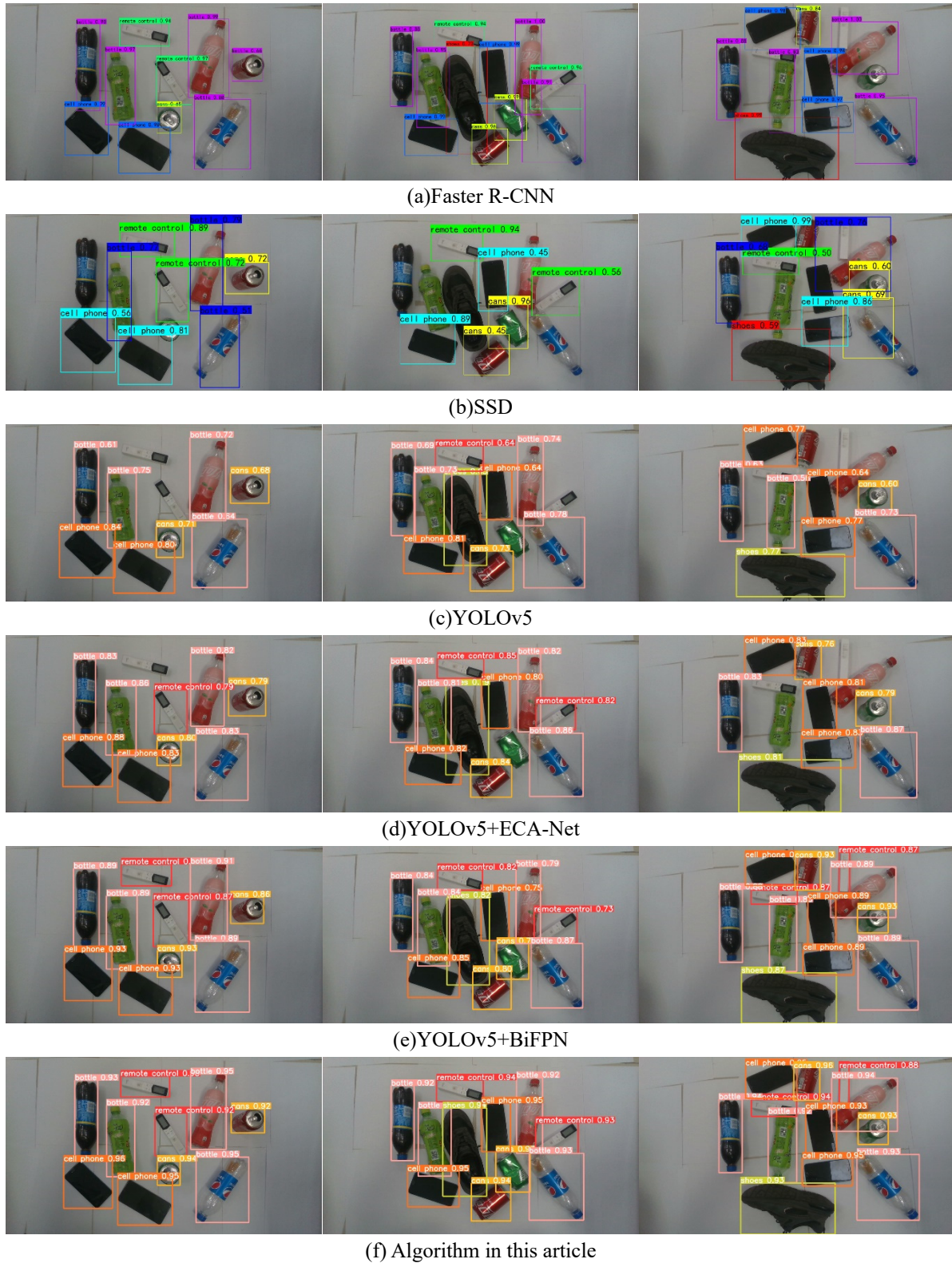


Figure 12. Comparison of Detection Results of Various Algorithms

There are a total of 6 types of algorithms in Figure 12, including Faster R-CNN, SSD, YOLOv5, YOLOv5+ECA-Net, YOLOv5+BiFPN, and the algorithm presented in this paper. The success rate of this experiment is defined as the percentage of correctly detected item categories to the total

number, and the average detection accuracy is the average of the three groups of experiments in the figure. (a) The success rates of the Faster R-CNN model detection in the figure are 90%, 100%, and 75%, respectively, with an average detection success rate of 88.33%. (b) The success rates of SSD model

detection in the figure are 80%, 54.54%, and 58.33%, respectively, with an average detection success rate of 64.29%. (c) The detection success rates of the YOLOv5 model in the figure are 80%, 81.81%, and 66.67%, respectively, with an average detection success rate of 76.16%. (d) The detection success rates of the YOLOv5+ECA-Net model in the figure is 90%, 90.90%, and 66.67%, respectively, with an average detection accuracy of 82.52%. (e) The figure shows the YOLOv5+BiFPN model with a detection accuracy and average detection accuracy of 100%. (f) The figure shows that the detection accuracy and average detection accuracy of the algorithm model in this article reach 100%, which has improved the accuracy of confidence compared to the suboptimal algorithm YOLOv5+BiFPN. Through the experimental results, it can be verified that the improvement of the algorithm in this paper is successful. During the detection process, it can reduce issues such as missed detections and false detections, greatly improving the detection accuracy of the algorithm.

5. Conclusion

With the development of object detection algorithms, high-precision and real-time detection technologies are urgently needed in daily life. The author has designed a detection algorithm based on daily necessities:

(1) By optimizing the clustering anchor box, the positioning accuracy of targets with different sizes is improved, and a suitable anchor box size is found for the targets.

(2) Replacing the original network's FPN+PAN structure with BiFPN structure facilitates the network to quickly fuse multi-scale features and improve the detection performance of targets at different scales.

(3) Adding the ECA Net attention mechanism after the BiFPN structure solves the problem of slow convergence in the training process due to complex backgrounds and large datasets.

Through the above improvements, the YOLOv5s algorithm was trained on the basis of a self-made dataset. The experimental results showed that the improved algorithm in this paper can provide detection accuracy for daily necessities, and the detection accuracy reaches 87.78%, which is 3.70% higher than the original YOLOv5.

References

- [1] X F Miao, B L Liu, X Q Li, et al. Improvement of YOLOv5s Rail Crack Target Detection Algorithm [J/OL]. Computer Engineering and Applications: 1-11 [2023-09-02].
- [2] P Y Wang, L Mao. Research on small target detection algorithm for lychee fruit based on YOLOv5 [J]. Shanxi Electronic Technology, 2023 (04): 74-77H. Poor, An Introduction to Signal Detection and Estimation. New York: Springer-Verlag, 1985, ch. 4.
- [3] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE Computer Society, 2014: 580.
- [4] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(6): 1137-1149.
- [5] Y Zhang, Z Y Gong, W W Wei. Traffic sign detection based on improved Faster R-CNN model [J]. Progress in Laser and Optoelectronics, 2020, 57(18): 173-181C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.
- [6] Y L Pei, H Luo, S H Zhang, et al. High speed rail fastener detection algorithm based on improved Faster R-CNN [J]. Journal of East China Jiaotong University, 2023, 40(01): 75-81M. Young, The Technical Writers Handbook. Mill Valley, CA: University Science, 1989.
- [7] REDMON J, FARHADI A. YOLOv3: an incremental improvement [J]. arXiv e-prints, 2018.
- [8] BOCHKOVSKIY A, WANG C Y, LIAO. YOLOv4: optimal speed and accuracy of object detection [R]. 2020.
- [9] X Y Niu, P G Mao, Y T Duan, et al. Research on Lightweight Improved Algorithm for Indoor Target Detection Based on YOLOv5s [J/OL]. Computer Engineering and Applications: 1-11 [2023-09-02].
- [10] G L Yang, J X Wang, Z L Nie. A real-time tomato recognition method based on improved YOLOv5s [J]. Jiangsu Agricultural Science, 2023, 51(15): 187-193.
- [11] C Zhang, Z Tian, J Song, et al. Construction worker hardhat-wearing detection based on an improved BiFPN [C]//2020 25th international conference on pattern recognition (ICPR). IEEE, 2021: 8600-8607.
- [12] Q Wang, B Wu, P Zhu, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11534-11542.
- [13] R Zhou, M Li, S G Meng, et al. Research on Airport Runway Intrusion Warning Technology Based on YOLOv5 and Deep SORT [J/OL]. Electronic Measurement Technology: 1-6 [2023-09-02].