Scientific Research Publishing

Open Journal of Applied Sciences, 2021, 11, 1287-1306
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

An Improved YOLOv3 Model for Asian Food Image Recognition and Detection

Xiaopei He[1], Dianhua Wang[2], Zhijian Qu[1*]

[1]School of Computer Science and Technology, Shandong University of Technology, Zibo, China
[2]Lianchi School of Zhangdian District, Zibo, China
Email: *1305838342@qq.com

How to cite this paper: He, X.P., Wang, D.H. and Qu, Z.J. (2021) An Improved YOLOv3 Model for Asian Food Image Recognition and Detection. *Open Journal of Applied Sciences*, **11**, 1287-1306.
https://doi.org/10.4236/ojapps.2021.1112098

Received: November 30, 2021
Accepted: December 27, 2021
Published: December 30, 2021

Abstract

The detection and recognition of food pictures has become an emerging application field of computer vision. However, due to the small differences between the categories of food pictures and the large differences within the categories, there are problems such as missed inspections and false inspections in the detection and recognition process. Aiming at the existing problems, an improved YOLOv3 model of Asian food detection method is proposed. Firstly, increase the top-down fusion path to form a circular fusion, making full use of shallow and deep features. Secondly, introduce the convolution residual module to replace the ordinary convolution layer to increase the gradient correlation and non-linearity of the network. Thirdly, introduce the CBAM (Convolutional Block Attention Module) attention mechanism to improve the network's ability to extract effective features. Finally, CIOU (Complete-IoU) loss is used to improve the convergence efficiency of the model. Experimental results show that the proposed improved model achieves better detection results on the Asian food UECFOOD100 data set.

Keywords

Asian Food, YOLOv3, Feature Fusion, Complete-IOU, CBAM

1. Introduction

With the development of science and technology and the improvement of human living standards, food object detection plays an important role in the fields of digital retail services, smart homes, healthy eating, self-eating detection, etc. More and more research based on food images analysis is emerging [1] [2].

In recent years, the classification and detection of food images has attracted widespread attention from scholars. Most researchers start from the structural

characteristics of food and base their research on the characteristics of the food itself. In 2009, Taichi Joutou [3] proposed an automatic food image recognition system, which uses Multiple Kernel Learning (MKL) method to integrate multiple image features to classify food images. In 2015, Bettadapura [4] used the food pictures taken by the camera and other information such as the location of the restaurant involved in the pictures to classify the food through support vector machines (SVM). However, traditional machine learning methods are less robust in real scenarios. With the development and application of deep learning models, the introduction of deep learning models into food image target detection has become the mainstream method. In 2016, Jingjing Chen [5] used the mutual and fuzzy relationship between foods and proposed a deep network structure that simultaneously learns food component recognition and food classification. In 2018, Eduardo Aguilar [6] focused on the cafeteria environment and carried out research on automatic food analysis, integrating multiple functions such as food positioning, recognition, and segmentation. In 2019, Zhang Gang and others [7] proposed a food image recognition method based on diffusion graph convolutional network and transfer learning, Weiqing Min [8] and others used abundant raw material composition information to locate multiple food images of different scales, and realized the recognition from the category level to the composition level. In 2020, Ya Lu [9] propose a novel system based on artificial intelligence (AI) to accurately estimate nutrient intake, by simply processing RGB Depth (RGB-D) image pairs captured before and after meal consumption.

Since the texture and color information contained in the food itself is too rich, it is confusing to the model. Therefore, food object detection is one of the most challenging tasks in the field of machine learning. In addition, the main source of food pictures is the dining table scene. In the pictures, there are inevitably background information such as tableware, other dishes, and sundries, which can easily affect the detection effect. As far as Asian food is concerned, its shape and structure are diverse, and the appearance of food under different cooking methods is also very different. These factors significantly increase the difficulty of detection. Moreover, using too much additional information such as raw materials and geographic location will lead to slow detection speed and poor real-time performance. When additional information is unavailable or unavailable, the accuracy of food detection will be greatly affected.

In order to solve the above-mentioned problems in the detection of Asian food targets, further explore the correlation between the inherent characteristics of Asian food and the detection results, and improve the accuracy of Asian food detectors, this paper improves the YOLOv3 [10] algorithm and uses the improved algorithm in Asia Food target detection. It aims to improve the accuracy of Asian food target detection under complex background conditions. The improved network structure is shown in **Figure 1**. Firstly, improve the feature fusion method. On the basis of the FPN bottom-up fusion path, add a top-down

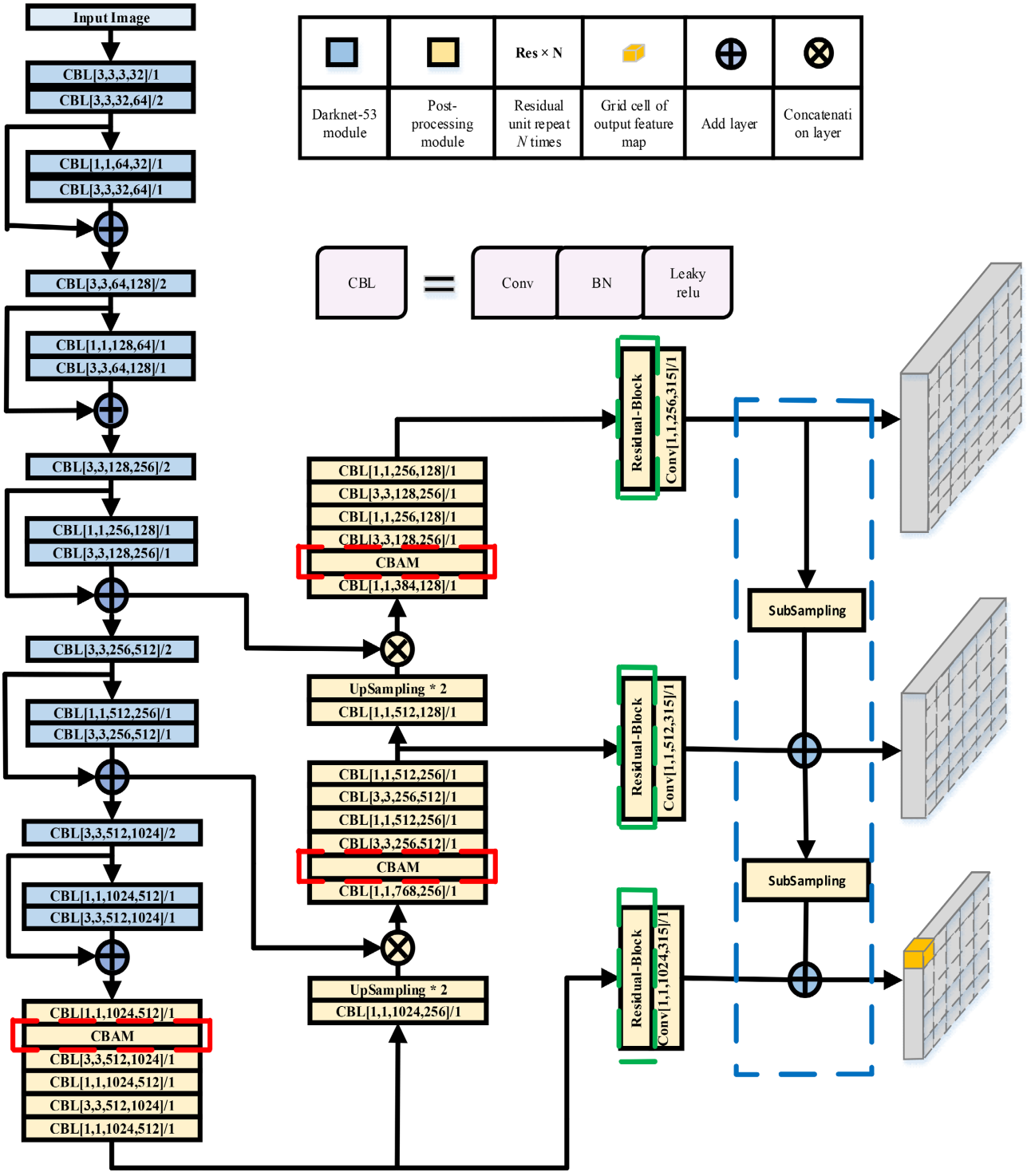


Figure 1. Improved structure of YOLOv3.

fusion path to form a ring fusion, and make full use of shallow and deep features. Secondly, introduce the convolution residual module [11] to reshape the feature output layer, effectively increasing the gradient correlation and nonlinearity of the network. Thirdly, introduce the CBAM (Convolutional Block Attention Module) attention mechanism [12] to improve the network's ability to

extract effective features. Finally, the CIOU (Complete-IoU) positioning loss function is used to fully consider the integrity of the food and improve the convergence efficiency of the model.

2. Improve YOLOv3 Model

YOLOv3 has high detection accuracy and meets real-time requirements. It is one of the target detection methods being used in a high frequency. YOLOv3 is divided into three parts: the feature extraction network DarkNet-53, Feature Pyramid Networks (FPN) [13] and the detection layer. Original YOLOv3 only transmits semantic information when performing FPN feature fusion, ignoring location information. In addition, when using the Mean Squared Error (MSE) regression loss function in the original YOLOv3, the deviation between the predicted box and the true box is relatively large. On the other hand, judging from the characteristics of the data itself, Asian food has richer textures and colors, which makes the network extract fewer discriminative features. Aiming at the existing problems, an improved YOLOv3 Asian food target detection network is proposed.

2.1. Annulus-FPN

The original YOLOv3 algorithm uses the FPN feature fusion network, which combines the feature information of three different scales of 52 × 52, 26 × 26, and 13 × 13 through the feature pyramid structure. The shallow network provides the location information of the target, and the deep network provides the semantic information of the target. After the feature fusion of FPN, different layers correspond to target detection of different scales. The shallow layer is used to detect small targets, and the deep layer is used to detect large targets. However, the FPN in the original YOLOv3 algorithm only transmits the semantic information of the deep features from top to bottom, and does not transmit the shallow position information. That different feature layers will have certain semantic expression and position expression for targets of different scales is ignored, which affects the accuracy of Asian food target detection.

In response to the above problems, this paper redesigned the feature fusion method of FPN, which is called "Annulus-FPN", as shown in **Figure 2**.

The red frame in **Figure 2** is the FPN feature fusion network structure, and the blue frame is the improved Annulus-FPN feature fusion network structure. The food pictures are extracted through the DarkNet-53 network, and the extracted feature maps are input to the FPN feature fusion network for bottom-up feature fusion. The Annulus-FPN feature fusion adds a top-down fusion path based on the FPN feature fusion, and transfers the precise position information of the shallow features to the deep features.

Feature fusion in FPN is different from that of Annulus-FPN. In FPN, the Concat operation is used to fuse the semantic information of the upper and lower features and splice in the channel dimension. The fusion method selected for

Annulus-FPN feature fusion is the add operation, which can increase the amount of information in each dimension of the fusion image, while the dimensionality of the image remains unchanged, so that the feature fusion network can use fewer parameters and achieve a higher computational efficiency. By enhancing the fusion path, the network can deliver more reliable semantic information and location information, and further utilize feature information to achieve feature enhancement.

2.2. The Convolutional Block

In the YOLOv3 network structure, in order to further extract features and obtain the prediction results of the network, the detection layer obtained after FPN feature fusion will perform 3 × 3 convolution and 1 × 1 convolution operations, of which the 3 × 3 convolution operation is shown in Figure 3.

Generally, the greater the degree of non-linearity of the convolutional neural network, the better the performance of the network. The degree of non-linearity of the network is closely related to the use of the activation function. Ordinary 3 × 3 convolution uses only one Rectified Linear Unit (Relu) activation function, and its degree of non-linearity is far from enough. When information is transferred, there will be more or less problems such as information loss. In addition, although the use of ordinary 3 × 3 convolution can deepen the network and help extract feature information, it will also cause the correlation between the gradients of the backpropagation to become worse.

In response to the above problems, the convolutional block is introduced to replace the ordinary 3 × 3 convolution output feature, and the convolutional block structure used is shown in Figure 4.

The convolution residual structure includes 1 × 1 convolution [14] [15], 3 × 3

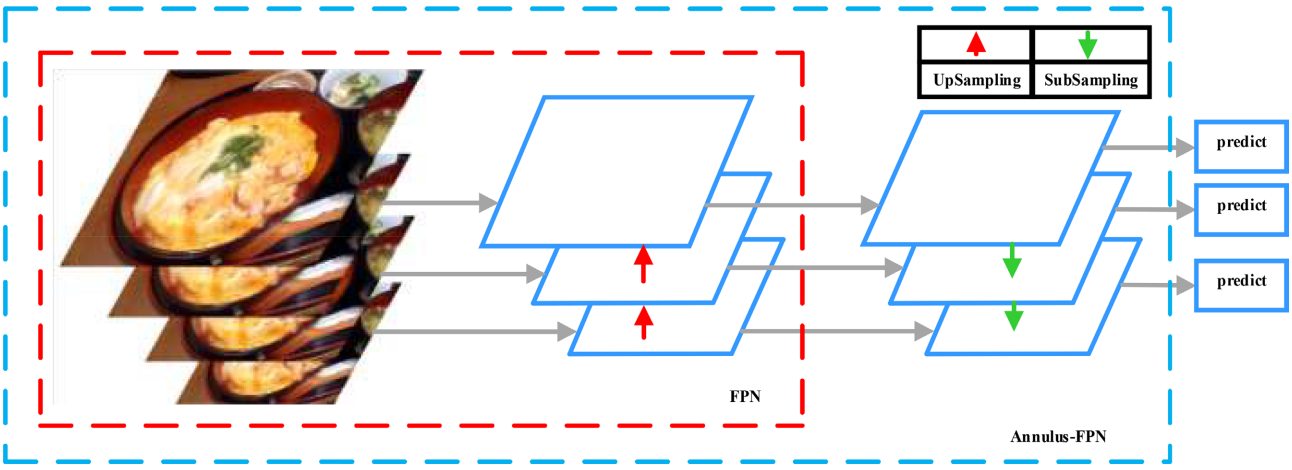


Figure 2. Annulus-FPN structure.

Figure 3. 3 × 3 Convolution.

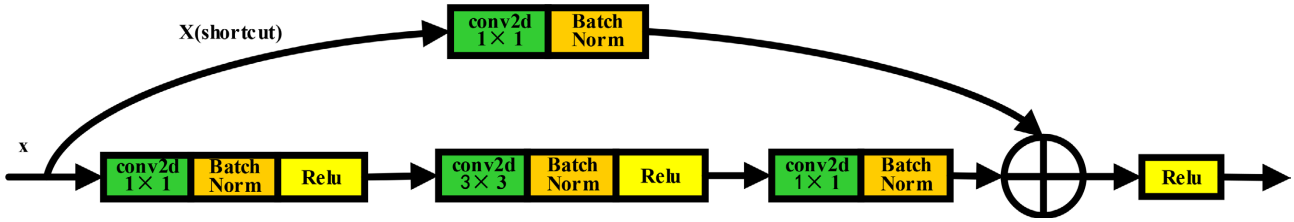


Figure 4. The convolutional block.

convolution, 1 × 1 convolution, shortcut, and addition operations. Among them, the former 1 × 1 convolution and the later one can achieve dimensionality reduction and dimensionality increase, so that the linear combination of information between channels can be changed, and the information interaction between channels can be realized. Under the premise of keeping the scale of the feature map unchanged, the Relu activation function can achieve large Increase the non-linear characteristics and improve the network performance. The 1 × 1 convolution in the Shortcut path is used to adjust the size of the input x in order to match the size of the feature map of the subsequent addition operation. To a certain extent, it solves the problem of information loss when the ordinary convolution is transmitted, and the input information is passed. The shortcut is passed to the output to protect the integrity of the feature information. In addition, the introduction of the convolution block can ensure that the back propagation of a gradient is the same as the forward propagation, which not only deepens the network but also maintains the gradient correla.

2.3. CBAM

Compared with other images, Asian food images have too rich color information and texture details. The features collected by the network are confusing, and there are fewer discriminative features about the true category, which leads to missed and false target detections. Therefore, the feature extraction of key areas of Asian food is particularly important.

The CBAM attention mechanism is a lightweight attention mechanism that integrates two dimensions of space and channel. The structure is shown in Figure 5. It infers the attention map in turn along these two independent dimensions, and then multiplies the attention map with the feature map to perform adaptive feature optimization.

CBAM contains CAM module and SAM module. CBAM uses the output matrix of the original neural network convolutional layer as the input matrix of the module. The CAM module simultaneously performs maximum pooling and average pooling operations on the input matrix in the channel dimension to obtain a one-dimensional channel attention matrix, which gives weights to the food characteristics of each channel. The SAM module performs global maximum pooling and average pooling on the front output matrix in the spatial dimension to obtain a two-dimensional spatial attention matrix. By embedding CBAM based on the channel and spatial attention mechanism in YOLOv3, the receptive

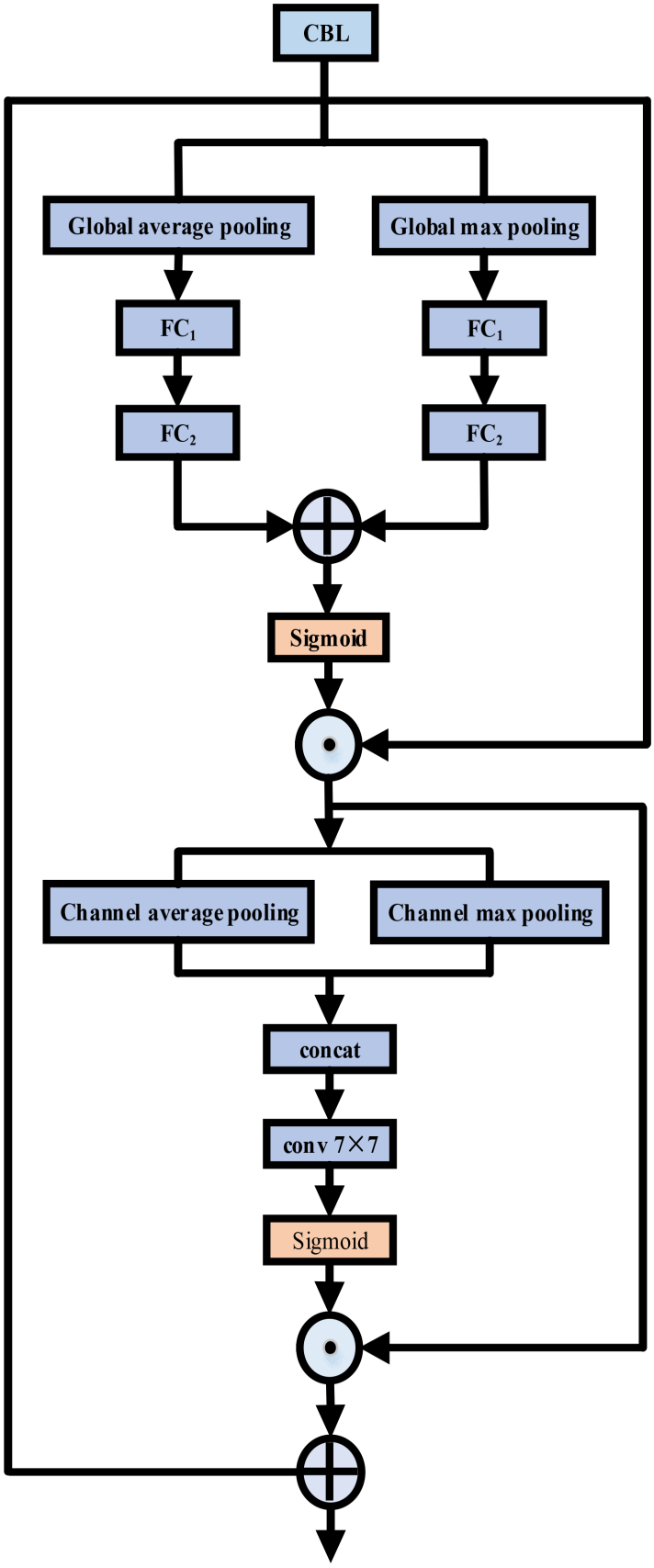


Figure 5. CBAM structure.

field of the network feature extraction layer is enhanced, and the feature extraction ability of the network is improved.

2.4. Regression Loss Function-CIoU

In addition, on the basis of the above improvements to the YOLOv3 model, in view of the large deviation between the predicted box and the real box, the CIoU regression loss function [16] [17] is introduced to replace the L2 norm loss function, which effectively improves the positioning accuracy of the bounding box and speed up the convergence of the model.

The loss function of the YOLOv3 network model is composed of three parts: target box position loss, target confidence loss and target category loss. Cross-entropy cost function (CE) is used to calculate the confidence loss and classification loss. The MSE which is also called the L2 norm loss function is used to calculate the regression loss of the bounding box position coordinates. However, there are some problems in using the MSE loss function to calculate the regression loss. First, there is a certain correlation between the center point and the width and height of the target box, but the MSE loss function regards the center point coordinates and width and height of the prediction box as independent variables, without considering the integrity of the target. Secondly, the L2 norm loss is sensitive to the scale of the bounding box, and the smaller the scale, the greater the influence of the bounding box prediction bias on it. Finally, the convolutional neural network uses Intersection over Union (IoU) as the standard when evaluating the regression effect of the bounding box, however, the optimization between L2 norm loss and IoU is not equivalent, and cannot accurately reflect the overlap between the truth box and the prediction box.

In response to these problems, CIoU is introduced instead of MSE as the model bounding box regression loss function. The CIoU loss function considers the overlap area, center distance and aspect ratio in the bounding box regression, and its loss function is defined as shown in Equation (1).

$$L_{coor} = 1 - \mathrm{IoU} + \frac{d_c^2\left(b^{\mathrm{p}}, b^{\mathrm{g}}\right)}{d_e^2} + av \tag{1}$$

Among them, L_{coor} represents the CIoU loss used for positioning loss, b^{p} and b^{g} represent the center points of the predicted bounding box and the true bounding box, respectively, and $d_c\left(b^{\mathrm{p}}, b^{\mathrm{g}}\right)$ is the distance between b^{p} and b^{g}, and d_e is the diagonal length of the smallest closed box covering the two boxes. av is a penalty factor, which can control the width and height of the prediction box and quickly fit the width and height of the real box. Among them, a represents the parameter of trade-off, and the definition is shown in Equation (2).

$$a = \frac{v}{\left(1 - \mathrm{IoU}\right) + v} \tag{2}$$

v represents a parameter for the consistency of the aspect ratio, and is defined

as shown Equation (3).

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^g}{h^g} - \arctan\frac{w^p}{h^p}\right)^2 \tag{3}$$

Among them, w^g and h^g are the width and height of the real box, w^p and h^p are the width and height of the prediction box, and a schematic diagram of the CIoU loss function is shown in Figure 6.

CIoU loss directly minimizes the normalized distance between two boxes, and takes into account of the three geometric properties: overlap area, center distance and aspect ratio, which can improve the stability of box regression and the accuracy of model convergence.

3. Experimental Data Set

3.1. Introduction to the Data Set

The data set uses the UEC-FOOD100 public Asian food data set. The data set has a total of 12,741 pictures, including 100 Asian foods such as rice, grilled chicken, and sweet and sour pork. The characteristics of the UEC-FOOD100 data set are small differences between food images, large differences within categories, many cooking methods, and rich information such as colors and textures. These can accurately reflect the complex and diverse characteristics of

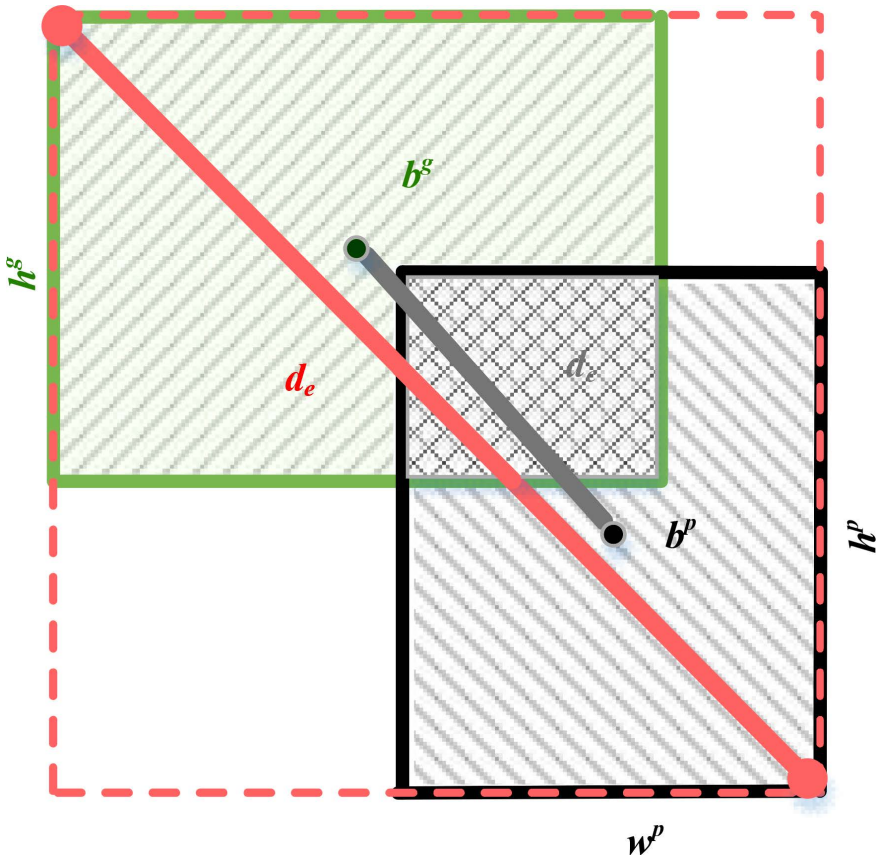


Figure 6. CIoU loss function.

Asian food.

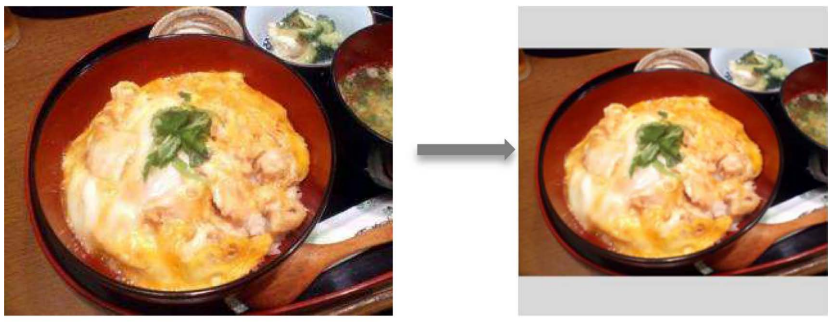
During the experiment, the UECFOOD100 data set was made into VOC2007 format. The marked center point, width and height, target object category and other information were saved as xml files required by VOC format. The data format in the UECFOOD100 data set is that each picture has one or more foods, and if there are multiple foods, it corresponds to multiple xml files. In the experiment, multiple xml files of each picture are merged, so that each picture corresponds to only one xml file, making the format is more concise, and the training speed faster. The data set is randomly divided into training set, validation set and test set, the proportions are 80%, 10% and 10%.

3.2. Data Enhancement

Through the analysis of the UECFOOD100 data set, the resolution of a large number of pictures is about 800 × 600 pixels, and the aspect ratio is about 4:3. When the data being input into the 416 × 416 YOLOv3 network, because the network adopts the resize processing method of maintaining the aspect ratio of the input picture, so the vacant part is filled with the pure gray pixels, as shown in Figure 7. The image whose original image size is 800 × 600 actually occupies only 416 × 312 in the network input, and the rest are pure gray pixels.

After the resize operation, the target zoom is too small, and the detection is difficult. About one-third of the network calculation is used to calculate the supplementary gray pixels, which causes a lot of waste of computing resources. In addition, the pictures in the data set are distributed in a long-tailed manner. There are as many as 723 pictures in each category, and as few as 100 pictures, as shown in Figure 8. For the target detection task, only more than 100 pictures in the category are used to train the deep learning model, and the amount of data is not sufficient.

In order to increase the amount of data and improve the effect of model training, Mosaic data enhancement [18] is used to simulate more data samples during the experiment. The food image after Mosaic data enhancement is shown in Figure 9. It can be seen from Figure 9 that the picture after data enhancement does not have a solid-color border. During the cropping process, the target



(a) (b)

Figure 7. Comparison of the original picture and the zoomed picture. (a) 800 × 600; (b) 416 × 416.

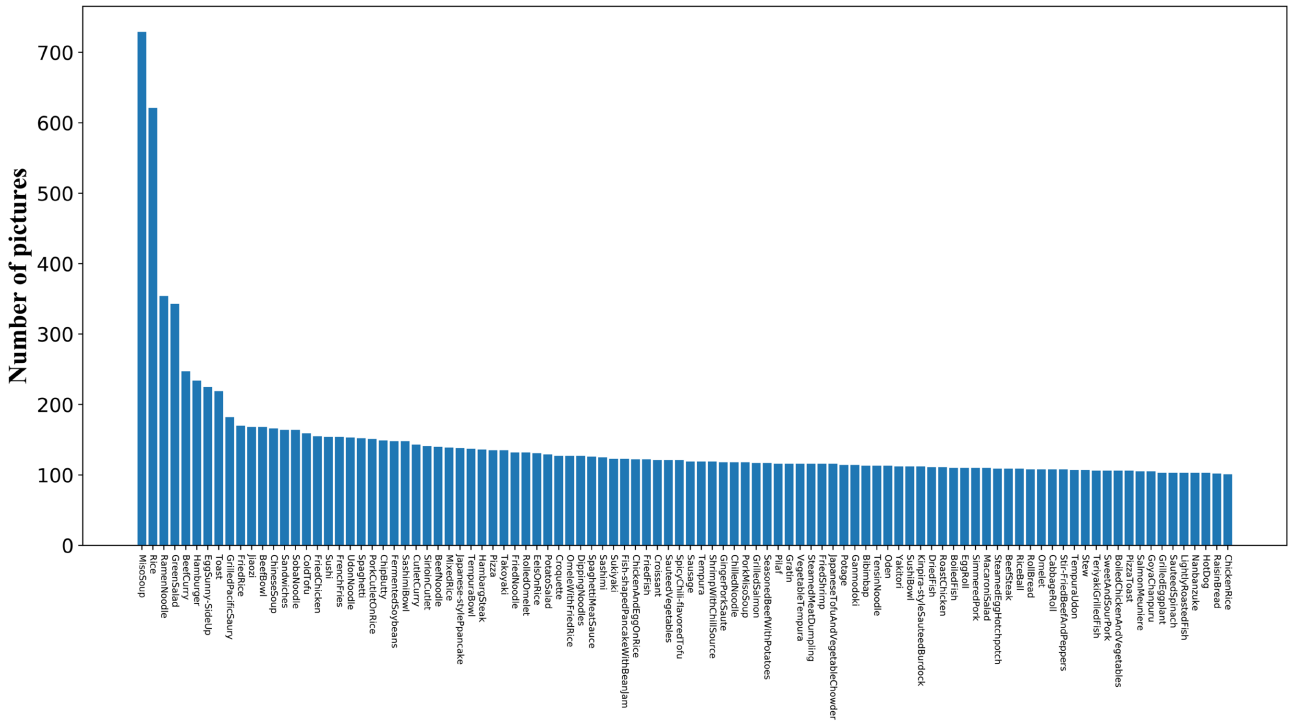


Figure 8. Data distribution in UECFOOD100 data set.

Figure 9. Picture enhanced with Mosaic data.

will not increase the difficulty of detection because it is too small, and it can simulate more sample data.

4. Experimental Results and Analysis

4.1. Experimental Environment Configuration and Model Training

Use Intel (R) Core (TM) i5-8500 CPU processor, NVIDIA GeForce GTX 1660 graphics card, 16 GB CPU memory, 6 G graphics card memory. Using the deep learning framework based on pytorch, the original YOLOv3 network and the improved YOLOv3 network were trained and analyzed separately under the windows operating system.

First, the UECFOOD100 Asian food data set was used to train the YOLOv3 pre-training model. In the training process, the learning rate dynamic adjustment strategy is adopted, and the step mode is selected to update the learning rate. At the beginning of training, due to the limited Graphics Processing Unit (GPU) memory, the initial learning rate is set to 1e−3, the gamma coefficient is set to 0.92, and the batch size is set to 8. In the 50th iteration, the model no longer converges. At this time, the learning rate is set to 1e−4, and the batch size is set to 4 to continue training. As the number of training increases, the loss value continues to decrease, and the loss value decline curve is shown in Figure 10. In Figure 10, the training loss is measured during each epoch, and the verification loss is measured after each epoch using the data in the verification set. According to the analysis of the convergence curve of the loss, the iteration is stopped after 100 rounds of training. It should be noted that in the 50th iteration, there are obvious fluctuations in the curve due to the change of the learning rate and batch size.

4.2. Evaluation Index

In the experiment, Average Precision (AP) and Mean Average Precision (mAP) [19] two common performance indicators in target detection algorithms are used to evaluate the target detection performance of different models. The calculation of AP requires the introduction of Precision and Recall. The Precision and Recall (P-R) curve can be obtained by taking Recall as the horizontal axis and Precision as the vertical axis. The area under the P-R curve is defined as AP, and the integral is used for calculation, as shown in Equation (4).

$$\mathrm{AP} = \int_0^1 p(r)\mathrm{d}r \tag{4}$$

The calculation of the above AP value is only for one category, and mAP can

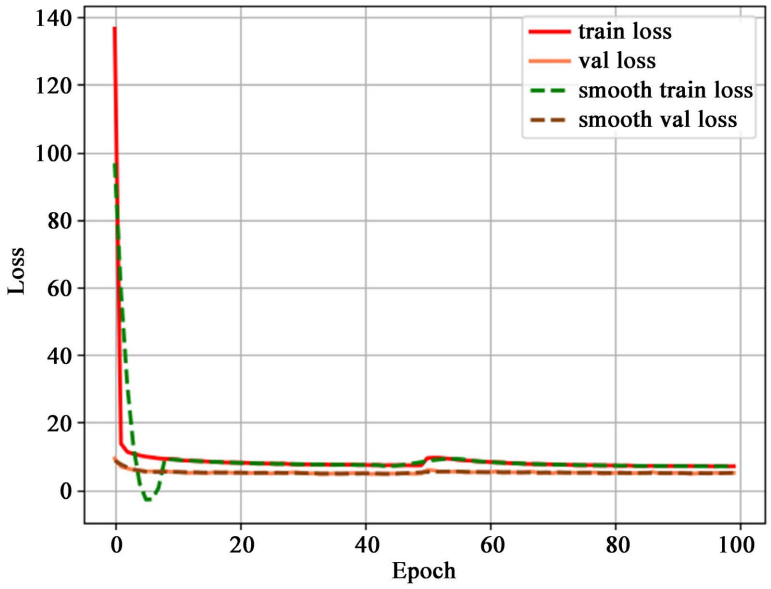


Figure 10. Training loss value curve.

be obtained by averaging the APs of all categories. mAP can measure the detection ability of the trained model in all categories. Assuming there are K categories, and $K > 1$, the mAP calculation method is shown in Equation (5).

$$\text{mAP} = \frac{\sum_{i=1}^{K} \text{AP}_i}{K} \tag{5}$$

4.3. Initial Assessment

After training on a 416 × 416 image, the performance of the original YOLOv3 was evaluated on the test set. Combining the recall rate and accuracy rate to evaluate the training network, consider the evaluation indicators under the two conditions of IoU = 0.5 (mAP@.5) and IoU = 0.75 (mAP@.75). The threshold of the confidence score (confidence score × category probability) of a specific category is set to 0.3 to generate the predicted bounding box. Table 1 summarizes the mAP@.5 and mAP@.75 indicators of the original YOLOv3.

In Table 1, Experiment 1 uses the original data set for training, and Experiment 2 adds Mosaic data enhancement for training. Under the indicators of IoU = 0.5 and IoU = 0.75, the test accuracy has been significantly improved.

4.4. Experimental Results and Analysis

In order to evaluate the proposed improvement method, the following four experiments were carried out: 1) YOLOv3-AFPN: The top-down fusion path is added to the original YOLOv3 FPN to form a annulus feature fusion. 2) YOLOv3-AFPN-RB: based on YOLOv3-AFPN, the convolutional block is introduced to replace the output features of the ordinary convolution layer. 3) YOLOv3-AFPN-RB-CS: based on YOLOv3-AFPN-RB,CBAM attention me-chanism is added. 4) YOLOv3-AFPN-RB-CS-CL: based on YOLOv3-AFPN-RB-CS, CIoU loss is used to replace the L2 norm loss of bounding box regression in the original YOLOv3. These four models were studied on the same training set, validation set and test set, and they all used the Mosaic data enhancement method.

It can be seen from Table 2 that under the index of IoU = 0.5, compared with the original YOLOv3 algorithm (Experiment 2 in Table 1) using the Mosaic data enhancement method, experiment 1 is to improve the feature fusion network to annulus feature fusion network and mAP is increased by 1.43%, indicating that the introduction of annulus feature fusion can make better use of deep and shallow features, and effectively improve the detection accuracy. Experiment 2 uses the convolutional block to replace the ordinary 3 × 3 convolution output features, mAP is improved by 0.94%, and the correlation of the gradient is effectively

Table 1. Test results of original YOLOv3.

Index	Algorithm	Training size	mAP_{50}/%	mAP_{75}/%
1	YOLOv3	416 × 416	70.54	50.49
2	YOLOv3 + Mosaic	416 × 416	72.58	51.91

Table 2. Comparison of experimental results after different improvements on the original YOLOv3 algorithm.

Index	Algorithm	Training size	mAP_{50}/%	mAP_{75}/%
1	YOLOv3-AFPN	416 × 416	74.01	52.98
2	YOLOv3-AFPN-RB	416 × 416	74.95	53.69
3	YOLOv3-AFPN-RB-CS	416 × 416	76.23	54.38
4	YOLOv3-AFPN-RB-CS-CL	416 × 416	77.60	55.52

maintained. Experiment 3 introduces the CBAM attention mechanism, and mAP is increased by 1.28%. It can be seen that CBAM can make the network pay more attention to discriminative features and improve the detection accuracy. Experiment 4 introduced the CIoU loss function instead of the IoU loss function, and the average accuracy increased by 1.37%, which means that CIoU played a more accurate role in the detection of bounding box regression.

Under the index of IoU = 0.5, the average accuracy of the original YOLOv3 algorithm is 70.54%, and the average accuracy of the improved YOLOV3 in this article has reached 77.60%. The change curve of the average accuracy of the model before and after the improvement in the training process is shown in Figure 11. The figure on the left is the average precision curve when IoU = 0.5, and the figure on the right is the average precision curve when IoU = 0.75. The above-mentioned average precision only reflects the overall level of the model, and cannot specifically reflect the performance of each category in the model. Therefore, this article draws a scatter plot of AP changes for each category before and after the improvement of YOLOv3, as shown in Figure 12, it can be seen that with the exception of individual categories, most categories have significantly improved AP after the improvement. The comparison of evaluation indicators between the improved YOLOv3 and the original YOLOv3 proves the effectiveness of methods such as Annulus-FPN, the convolutional block, CBAM attention mechanism, CIOU loss function, and data enhancement.

As shown in Figure 13, by outputting the heat maps (a) and (b) of the network before and after the improvement, it can be seen that the improved network is more targeted in terms of feature extraction and expands the target area. From Figure 13(c) and Figure 13(d), it can be seen that the improved network can effectively determine the type of food and improve the confidence of the network.

The original trained YOLOv3 network and the improved YOLOv3 network (YOLOv3-AFPN-RB-CS-CL) were used to detect the pictures in the test set. The detection results are shown in Figure 14. Among them, the left image in (a), (b), (c) is the detection result of the original YOLOv3, and the right image is the detection result of the improved YOLOv3. The network uses boxes to mark the size and location of the detected food target, and at the same time marks the classification and confidence of the target. From Figure 14(a), it can be found that the original YOLOv3 shows good effects on some targets, and the improved YOLOv3

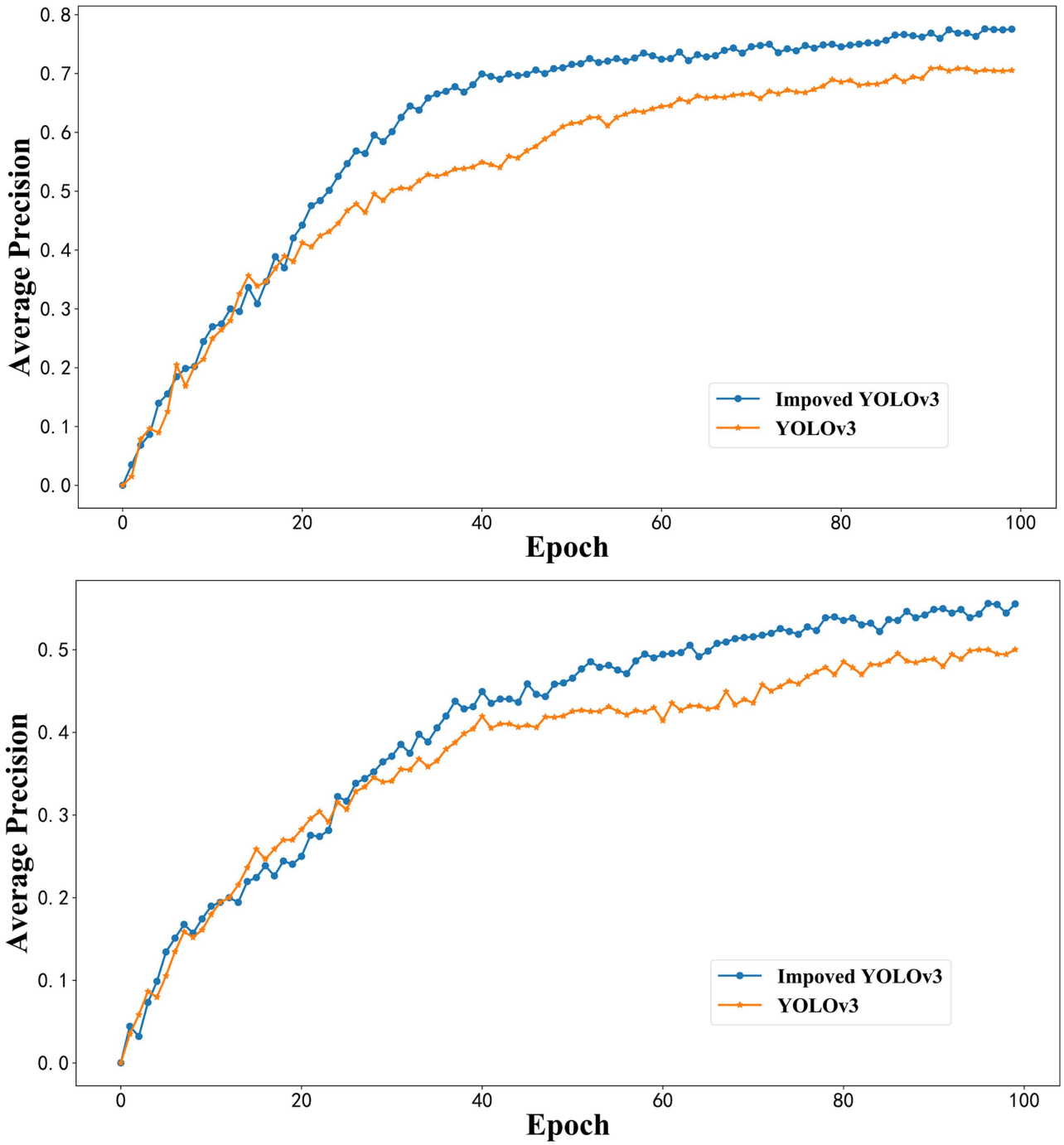


Figure 11. The average accuracy curve of different models in the training process.

also inherits the network can obtain a higher degree of confidence in a variety of food targets, with fewer false detections and missed detections. When compared with the original YOLOv3, it always shows better performance.

In order to further verify the effectiveness of the algorithm on the Asian food data set in this paper, the algorithm in this article is combined with 4 representative or advanced target detection algorithms SSD [20], Faster R-CNN [21], original YOLOv3, and EfficientDet-d2 [22] for comparison test, the performance

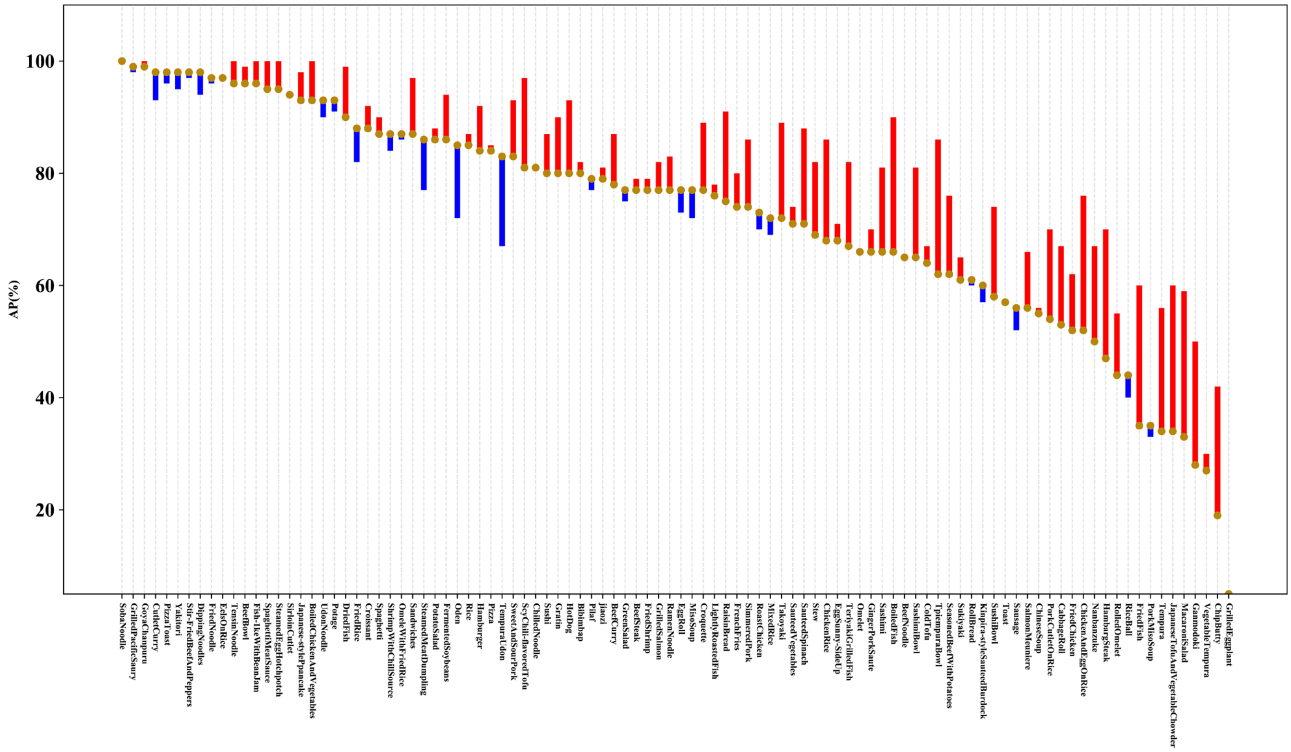


Figure 12. The YOLOv3 model before and after the improvement of the 100 types of food target detection AP value analysis in the data set (the red line represents the improved performance gain, the blue line represents the performance decline).

Figure 13. (a) Heat map before network improvement; (b) heat map after network improvement; (c) original image detection result before network improvement; (d) original image detection result after network improvement

comparison is shown in Table 3. It can be seen from Table 3 that under the index of IoU = 0.5, compared with other mainstream target detection algorithms, the algorithm in this paper has higher detection accuracy in Asian food target detection, and the mAP value reaches 77.60%, compared with the original YOLOv3 algorithm, the mAP value is increased by 7.06%.

The comparison of the detection results of the improved YOLOV3 algorithm in this paper with the original YOLOv3, good effects of the original YOLOv3.

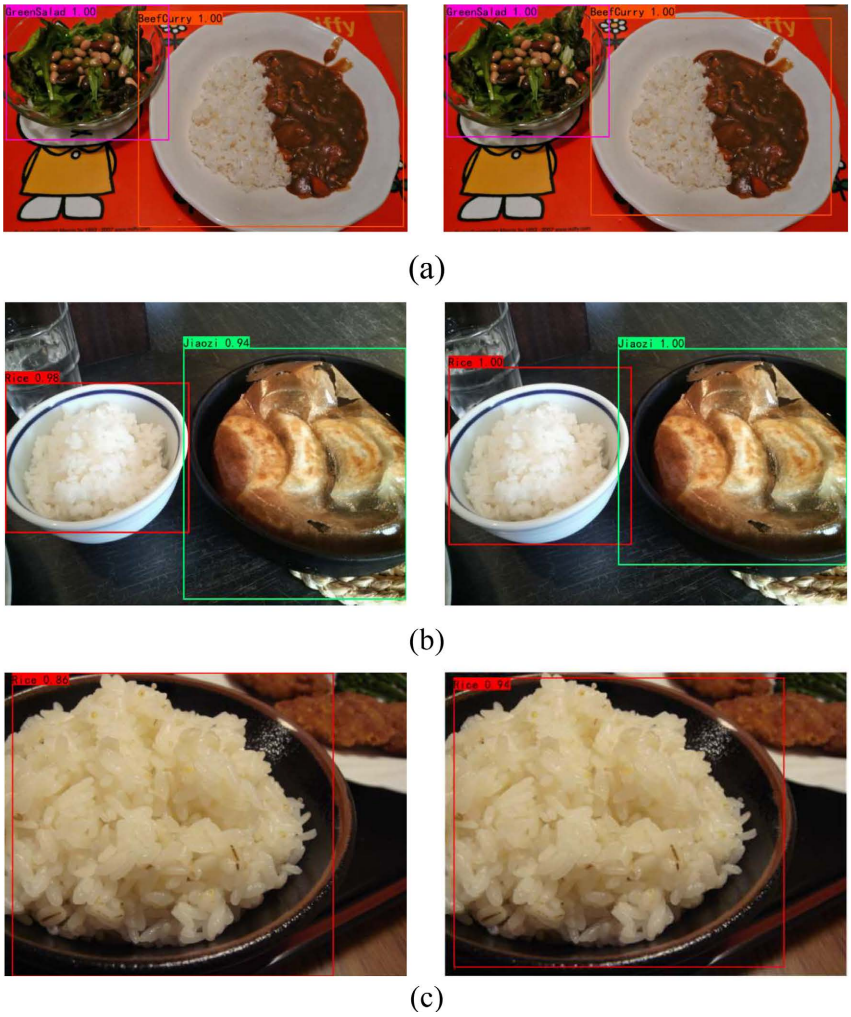


Figure 14. Comparison of detection results before and after improvement of YOLOv3.

Table 3. Performance comparison of mainstream algorithms.

Index	Algorithm	Training size	mAP_{50}/%	mAP_{75}/%
1	SSD	416 × 416	61.30	45.92
2	Faster R-CNN	416 × 416	65.38	48.50
3	EfficientDet-d2	416 × 416	69.56	50.13
4	YOLOv3	416 × 416	70.54	50.49
5	Improved YOLOv3	416 × 416	77.60	55.52

From Figure 14(b) and Figure 14(c), it can be found that the original YOLOv3 has low confidence in the detection of some targets, but the improved YOLOv3 SSD, Faster R-CNN, and EfficientDet algorithms is shown in Figure 15.

From the different performance of each model on the same image in Figure 15, it can be seen that SSD, Faster R-CNN, EfficientDet-d2, and original YOLOv3 have poor positioning accuracy, and the prediction confidence is generally low, and there is a problem of missed detection. The model in this paper has a

Figure 15. Comparison of detection effects of different algorithms (from top to bottom: SSD, Faster R-CNN, EfficientDet, YOLOv3, Improved YOLOv3).

significant improvement in the confidence of detecting the target and the positioning of the target is more accurate. In summary, compared with other algorithms, this algorithm is more suitable for Asian food target detection, and the detection effect has been significantly improved.

5. Conclusion

This paper improves the traditional YOLOv3 model based on the characteristics of Asian food pictures, and applies the improved model to Asian food pictures for target detection and recognition. By introducing a top-down fusion path in FPN feature fusion, and forming a ring-shaped feature fusion with the bottom-up fusion path in FPN, the reliability of model semantic information and location information transmission is improved, and food features are enhanced. The convolutional block is used in the output layer to replace the ordinary 3×3 convolution to maintain the gradient correlation and increase the degree of nonlinearity of the network. Through the CBAM attention mechanism and the CIoU loss function, the network's ability to express the characteristics of Asian food and the convergence speed of the model is improved. The results show that for Asian food target detection tasks, the performance of the improved YOLOv3 model has been significantly improved, and it also has greater advantages compared with other conventional target detection algorithms.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Mao, R., He, J., Shao, Z., *et al.* (2021) Visual Aware Hierarchy Based Food Recognition. *International Conference on Pattern Recognition*, Milan, 10-15 January 2021, 571-598.

[2] Lu, Y., Stathopoulou, T. and Mougiakakou, S. (2021) Partially Supervised Multi-Task Network for Single-View Dietary Assessment. 2020 25*th International Conference on Pattern Recognition* (*ICPR*), Milan, 10-15 January 2021, 8156-8163. https://doi.org/10.1109/ICPR48806.2021.9412339

[3] Joutou, T. and Yanai, K. (2009) A Food Image Recognition System with Multiple Kernel Learning. 2009 16*th IEEE International Conference on Image Processing* (*ICIP*), Cairo, 7-10 November 2009, 285-288. https://doi.org/10.1109/ICIP.2009.5413400

[4] Bettadapura, V., Thomaz, E., Parnami, A., *et al.* (2015) Leveraging Context to Support Automated Food Recognition in Restaurants. 2015 *IEEE Winter Conference on Applications of Computer Vision*, Waikoloa, 5-9 January 2015, 580-587. https://doi.org/10.1109/WACV.2015.83

[5] Chen, J.J. and Ngo, C.W. (2016) Deep-Based Ingredient Recognition for Cooking Recipe Retrieval. *Proceedings of the* 24*th ACM International Conference on Multimedia*, Amsterdam, 15-19 October 2016, 32-41. https://doi.org/10.1145/2964284.2964315

[6] Aguilar, E., Remeseiro, B., Bolaños, M., *et al.* (2018) Grab, Pay, and Eat: Semantic Food Detection for Smart Restaurants. *IEEE Transactions on Multimedia*, **20**, 3266-3275. https://doi.org/10.1109/TMM.2018.2831627

[7] Zhang, G. and Zhang, S.Q. (2019) Food Image Recognition Based on DCNN and Transfer Learning. *Laboratory Research and Exploration*, **38**, 111-114.

[8] Min, W.Q., Liu, L.H., Luo, Z.D., *et al.* (2019) Ingredient-Guided Cascaded Multi-Attention Network for Food Recognition. *Proceedings of the* 27*th ACM International Conference on Multimedia*, New York, 21-25 October 2019, 1331-1339. https://doi.org/10.1145/3343031.3350948

[9] Lu, Y., Stathopoulou, T., Vasiloglou, M.F., *et al.* (2020) An Artificial Intelligence-Based System to Assess Nutrient Intake for Hospitalised Patients. *IEEE Transactions on Multimedia*, **23**, 1136-1147. https://doi.org/10.1109/TMM.2020.2993948

[10] Redmon, J. and Farhadi, A. (2018) Yolov3: An Incremental Improvement. ArXiv: 1804.02767.

[11] He, K.M., Zhang, X.Y., Ren, S.Q., *et al.* (2016) Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[12] Woo, S., Park, J., Lee, J.Y., *et al.* (2018) CBAM: Convolutional Block Attention Module. *Proceedings of the European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[13] Lin, T.Y., Dollár, P., Girshick, R., *et al.* (2017) Feature Pyramid Networks for Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 936-944. https://doi.org/10.1109/CVPR.2017.106

[14] Lin, M., Chen, Q. and Yan, S. (2013) Network in Network. ArXiv: 1312.4400.

[15] Szegedy, C., Vanhoucke, V., Ioffe, S., *et al.* (2016) Rethinking the Inception Architecture for Computer Vision. *Proceedings of the IEEE Conference on Computer*

Vision and Pattern Recognition, Las Vegas, 27-30 June 2016, 2818-2826. https://doi.org/10.1109/CVPR.2016.308

[16] Zheng, Z.H., Wang, P., Liu, W., *et al.* (2020) Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 12993-13000.

[17] Zhang, L.D. and Deng, C. (2021) Multi-Scale Fusion of YOLOv3 Crowd Mask Wearing Detection Method. *Computer Engineering and Applications*, **57**, 283-290. http://kns.cnki.net/kcms/detail/11.2127.TP.20210602.1618.020.html

[18] Mu, S.Q., Lin, J.J., Wang, H.Q., *et al.* (2021) X-Ray Image Contraband Detection Algorithm Based on Improved YOLOv4. *Acta Armamentarii*. http://kns.cnki.net/kcms/detail/11.2176.TJ.20210922.0850.002.html

[19] Everingham, M., *et al.* (2010) The PASCAL Visual Object Classes (VOC) Challenge. *International Journal of Computer*, **88**, 303-338. https://doi.org/10.1007/s11263-009-0275-4

[20] Liu, W., Anguelov, D., Erhan, D., *et al.* (2016) SSD: Single Shot Multibox Detector. *European Conference on Computer Vision*, Amsterdam, 8-16 October 2016, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[21] Ren, S., He, K., Girshick, R., *et al.* (2015) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems*, **28**, 91-99.

[22] Tan, M.X., Pang, R.M. and Le, Q.V. (2020) EfficientDet: Scalable and Efficient Object Detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 10778-10787. https://doi.org/10.1109/CVPR42600.2020.01079