ISSN: 2414-1895 DOI: 10.6919/ICJE.202405_10(5).0038

# Optimisation and Realization of Improved Algorithm for Recognition of ChaoTian Pepper based on YOLOv5s

Xingke Zhang, Mingge Sun, Xiaolong Guo, Yinggang Li, Cong Gao

School of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin 132022, China

## Abstract

**In the process of recognising Chaotian peppers, due to the dense covering of the peppers, there is a leakage detection phenomenon and the recognition accuracy is not high. To this end, an improved YOLOv5s target recognition algorithm for Chaotian pepper is proposed. Firstly, the original 3×3 convolutional kernel is replaced with RepVGG module in the backbone network to enhance the extraction of feature information in complex and dense environments; Secondly,the introduction of an improved C3 module in the backbone network in conjunction with the CA(Channel Attention) attention module to enhance the ability to focus on important information, thus enhancing detection accuracy; Then by introducing the BiFPN(Bidirectional Feature Pyramid Network) module, the range of fused feature information is increased at multiple scales, which improves the model detection capability; On the Chaotian pepper dataset, a comparison experiment between the original YOLOv5s algorithm and the improved YOLOv5s algorithm shows The mAP(Mean Average Precision) value of the improved YOLOv5s algorithm is 3 percentage points higher than that of the original algorithm.There is also a large improvement in mAP in comparison with other mainstream algorithms.**

## Keywords

**Deep Learning; YOLOv5s; Chilli Pepper Recognition; Attention Mechanism.**

## 1. Introduction

China is the world's top producer and consumer of chilli peppers, and in recent years China's production of chilli peppers has reached more than 11 per cent of the country's total vegetable production[1], Among them, Guizhou has become the first province in China to grow chilli peppers, with the planting area accounting for one-sixth of the total planting area in China[2], Chaotian pepper is the largest variety of chilli pepper grown in Guizhou, and is the most important local agricultural product with great economic value. The current domestic picking of Chaotian pepper basically adopts artificial main, mechanical as a supplement[3]. Chao Tian Pepper is mostly ripe in summer, manual picking task is heavy, labour intensity is large. When picking mechanically, it is impossible to distinguish the maturity of the peppers and carry out indiscriminate picking, and there are also problems such as missed picking and excessive fruit damage, resulting in a large amount of wasted resources. Therefore, it is of great practical significance to study automatic picking robots, in which recognising morning glory peppers is the most important part of picking robots. Currently, target detection techniques are mainly classified into traditional target detection methods and deep learning based target detection methods. Traditional detection methods are feature extraction methods based on features such as colour[4], shape and texture of the target, which are highly influenced by human factors and have low recognition accuracy[5]. The deep learning technology, with its superb feature extraction capability and qualitative leap in both accuracy and speed, has gradually become the main means of target recognition. Common target detection models include Fast-RCNN[6], SSD[7], and

ISSN: 2414-1895 DOI: 10.6919/ICJE.202405_10(5).0038

YOLO[8]. Fengnan Shang et al. proposed an improved YOLOX algorithm to detect dragon fruits, which is conveniently deployed on embedded devices with a final accuracy of 98.9%[9]. Lu et al. proposed an improved YOLOv4 algorithm for the problem of small dense citrus fruits in orchard environments for fast fruit identification, maintaining real-time while improving accuracy[10]. Liu et al. deployed a lightweight improved YOLOv5 network on an unmanned aerial vehicle to achieve detection of apples on trees.

## 2. YOLOv5s Algorithm

YOLOv5 is an advanced target detection model developed by Ultralytics, the YOLOv5 model family includes YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x models of different scales[11]. The width, depth and number of parameters of the network gradually increased in the four versions, and although the detection accuracy also gradually increased but the detection speed also decreased. Among them, the YOLOv5s version is only a little less accurate than the other three versions but has the fastest detection speed and a much smaller number of parameters, which is most suitable for robotic picking scenarios. Because of its high performance, high accuracy, light weight and ease of use, YOLOv5 is widely used in a variety of practical scenarios.

The YOLOv5 network structure is shown in Fig. 1, including the input, the backbone part, the neck and the detection head part. The image input comes in and is first enhanced by Mosaic data, and the basic structure of Conv, C3, and SPPF is used in Backbone to extract features from the input image, Conv downsamples the input, C3 performs feature extraction and fusion, and SPPF uses pooling to enrich the semantic information of the features. Then the semantic information of the deep features and the location information of the shallow features are fully fused in Neck. Finally the processed feature map is predicted.

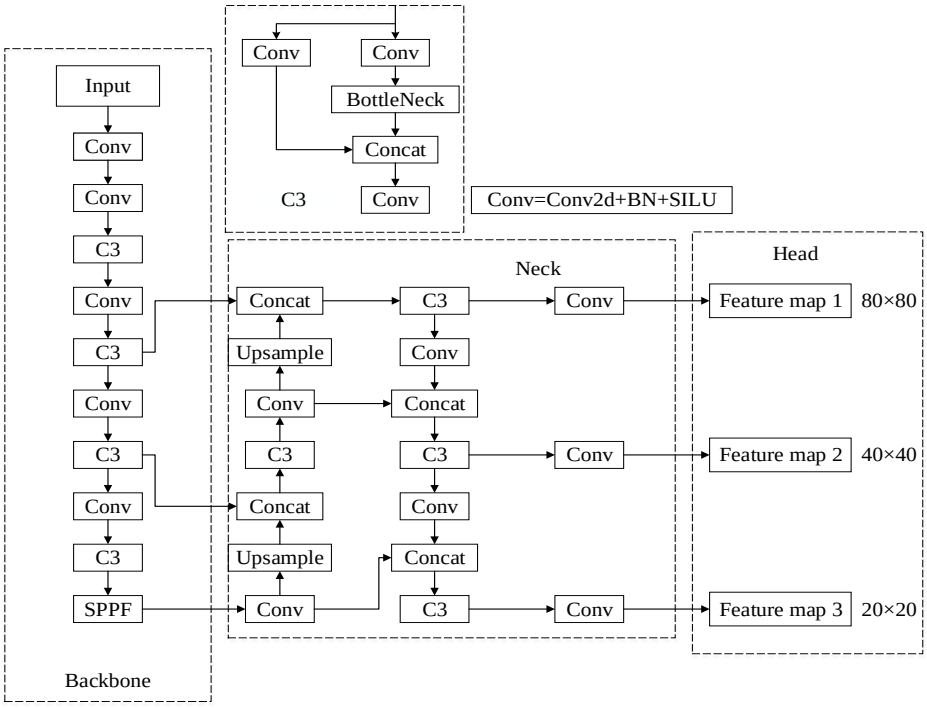



**Fig. 1** YOLOv5 network architecture

## 3. Improved YOLOv5s Algorithm

### 3.1 Overview of the Overall Improvement Algorithm

Aiming at the problem that chilli peppers grow densely and shade each other in complex natural environments, the YOLOv5 algorithm is improved as follows:

ISSN: 2414-1895 DOI: 10.6919/ICJE.202405_10(5).0038

(1) Use RepVGG module in Backbone to replace Conv to effectively improve the feature extraction capability;

(2) Add CA attention module in Backbone to fuse with C3 module and replace Bottleneck in C3 module, so that the model focuses on more important information to improve accuracy;

(3) Use BiFPN in Neck to enhance the fusion of multi-scale features; the improved YOLOv5 network structure is shown in Fig. 2.

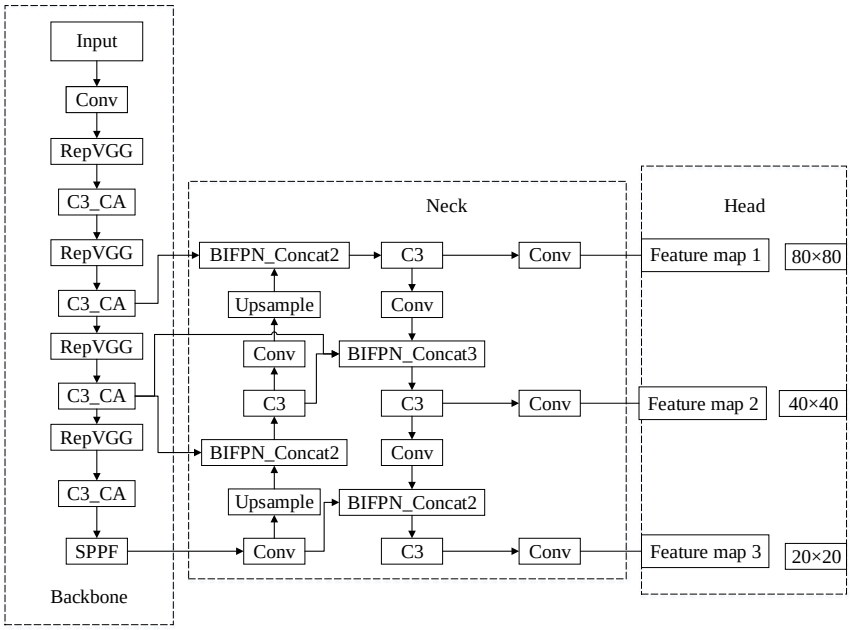



**Fig. 2** Improved YOLOv5 network structure

### 3.2 Improved Feature Extraction Backbone Network

The RepVGG module is a four-layer structure, including a residual structure consisting of a 3×3 convolution, a 1×1 convolution, and an Identity residual structure, the RepVGG structure is shown in Fig. 3.

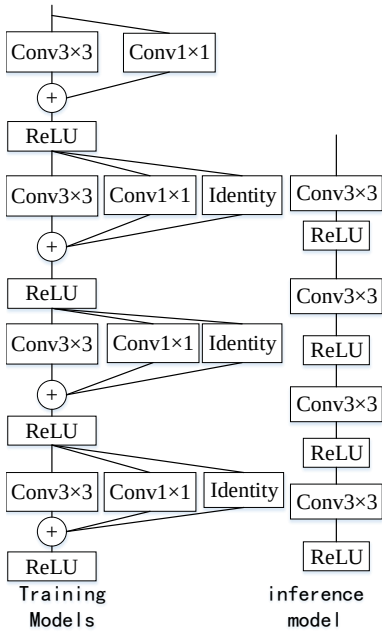



**Fig. 3** RepVGG structure

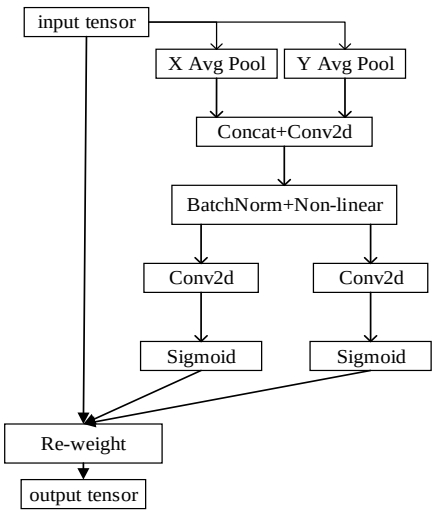
ISSN: 2414-1895 DOI: 10.6919/ICJE.202405_10(5).0038

Different network architectures are used for the network training phase and the network inference phase. The RepVGG network architecture for the training phase is a multi-branch structure with residual edges of 1×1 convolution added next to the 3×3 convolution in the first layer, while the remaining three layers also contain Identity residual structures. These residual structures effectively introduce multi-channel gradient propagation paths, which enables parallel training and mixing of multiple networks for better feature extraction to maximise their performance. Whereas the RepVGG network structure in the inference phase merges the branches to make a single branch structure, this structure consists of 3×3 convolution and Relu activation functions for easy model inference and acceleration. The number of parameters is increased but the accuracy is improved by using RepVGG module.

### 3.3 Adding the Attention Module

Attention is generally categorised into channel attention and spatial attention. The channel attention mechanism has a significant impact in improving model performance, however, location information is usually ignored. The introduction of CA attention embeds location information into channel attention, allowing better target localisation with little additional computational overhead and improved network accuracy. Firstly the input X is subjected to a coding operation using two 1D pooling kernels in the horizontal and vertical directions respectively. This operation uses one-dimensional pooling i.e., it avoids the loss of positional information due to two-dimensional pooling and preserves the positional information in the generated feature map. Next the encoding results are merged to obtain a new feature map by a simple stacking operation. The merged feature map is then subjected to a convolution operation to capture the relationship between the horizontal and vertical dimensions followed by the application of normalisation and activation functions to further process the features and obtain a richer feature representation. Subsequently, the new feature map was divided into two independent vectors applying 1×1 convolution to adjust the number of channels to adapt it to the attention computation. Using the Sigmoid activation function, the attention weights on the horizontal and vertical dimensions are obtained, which are used to represent the importance of different positions. Finally the original input feature map is multiplied with the attention weights on horizontal and vertical dimensions respectively to get the output of CA attention module.CA attention is shown in Fig. 4.

Using the C3 module fused with CA attention, the Bottleneck module in the C3 module is replaced with the CA attention module, and the model extracts the most important information, which improves the detection accuracy. The improved CAC3 module is shown in Fig. 5.



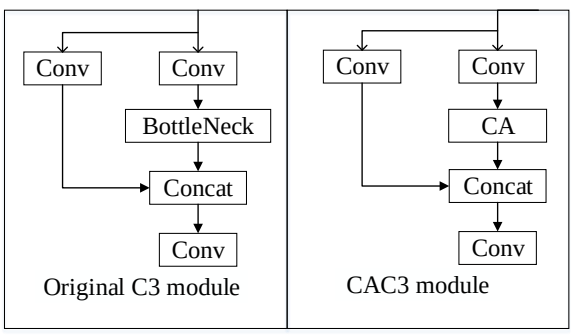




**Fig. 4** CA Attention Module **Fig. 5** Original C3 and Improved CAC3 Module Diagrams

ISSN: 2414-1895 DOI: 10.6919/ICJE.202405_10(5).0038

### 3.4 Improved Feature Fusion Network

Feature fusion is performed in the Neck section, where how to fuse shallow and deep information more efficiently is the key to improving the model. Most of the Chaotian pepper recognition are small targets and densely occluded, YOLOv5 algorithm detection is prone to the problem of missed detection, so BiFPN is used to improve the Neck network. The original YOLOv5 algorithm uses the PANet network structure, PANet has one bottom-up channel which is good for target localisation and another top-down channel which is good for target classification. The new Neck structure BiFPN is improved on PANet. As shown in Fig. 6, firstly the single-input node in the PANet is deleted, which effectively eliminates some redundant computations since this node has the same information as the previous node, and secondly an edge is connected between the original input node and the output node in the same layer for fusing more features. The most important of these is that traditional feature fusion is simply an overlay of feature maps, however, different input feature maps have different resolutions causing the output features to be affected by this. Thus BiFPN uses a weighted feature fusion approach to add an extra weight to each input. In short BiFPN is equivalent to enhanced PANet combined with weighted fusion technique.BiFPN is calculated as shown in equation (1).

$$O = \sum_i \frac{\omega_i}{\epsilon + \sum_j \omega_j} \cdot I_i \quad (1)$$

where ω is the learnable weights, ε is the very small value learning rate, I is the input value and O is the output value.

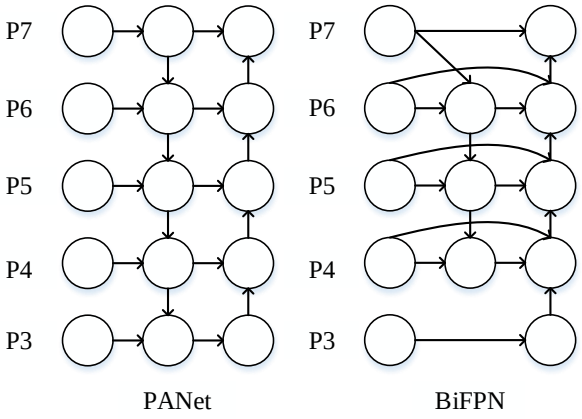



**Fig. 6** PANet structure and BiFPN structure

## 4. Experiments and Analysis of Results

### 4.1 Experimental Environment

The device operating system used in the experiment is ubuntu18.04, the CPU model is Intel(R) Xeon(R) Silver 4214R CPU @ 2.40GHz, and the GPU model is NVIDIA GeForce RTX 3080Ti. the experiment uses Python3.8 programming language to build the model and train the model through Pytorch1.8.1 The deep learning architecture is used to build and train the model, and Cuda11.1 architecture is used to accelerate the GPU.The training parameters are: the image size is 640×640, the Batch size is set to 16, the initial learning rate is 0.01, the momentum coefficient is 0.937, and the training period is 200 epochs.

### 4.2 Dataset

The Chaotian pepper dataset used for the experiments was obtained from Roboflow website, and an example of the images is shown in Fig.7. The dataset has a total of 2595 images, which contains images enhanced with data such as rotation, mirroring, exposure, Mosaic, etc. in order to enhance the complexity of the dataset and to expand the diversity of the samples, and they are divided into two

ISSN: 2414-1895 DOI: 10.6919/ICJE.202405_10(5).0038

categories: mature and immature. The label file format is in YOLO format and the dataset is divided into training set, validation set, and test set in the ratio of 7:2:1.

**Fig. 7** Partial sample of the dataset

### 4.3 Evaluation Metrics

The experiments use the evaluation indexes such as mAP (mean average precision), Params (number of parameters), GFLOPs (amount of computation), etc. to prove that each improvement point improves the performance of the model and outperforms the other mainstream algorithms. mAP calculation formula is shown in equation (2).

$$\mathrm{mAP} = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} \mathrm{AP}(i) \tag{2}$$

Where mAP is the average of all AP values, N is the number of categories, and AP is the average precision refers to the area enclosed by the curve composed of the horizontal coordinate of R (recall) and the vertical coordinate of P (precision), as shown in the formula (3) (4) (5).

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{3}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{4}$$

$$\mathrm{AP} = \int_0^1 \mathrm{PdR} \tag{5}$$

Where: tp refers to positive samples that were correctly classified; fp refers to negative samples that were incorrectly classified; fn refers to positive samples that were incorrectly classified; and mAP is an important measure of how well the model performs.

### 4.4 Ablation Experiments and Analyses

In order to verify that each step of improvement is effective, several ablation experiments were done. BiFPN, CA attention, and RepVGG were added to the YOLOv5 model.The experimental results are shown in Table 1, and "√" indicates that the module was added. The mAP increased by 1.7% after adding Improvement 1, 0.9% after adding Improvement 2, and 0.4% after adding Improvement 3. By adding these three improvements to the model together, the number of parameters and computation amount only slightly increased but compared to the original YOLOv5 algorithm, the mAP increased

ISSN: 2414-1895 DOI: 10.6919/ICJE.202405_10(5).0038

by 3%, thus only reducing the detection speed a little bit but effectively improving the model performance to increase the detection accuracy and enhance the recognition effect of dense artichoke peppers.

**Table 1.** Results of ablation experiments

| methodology | BiFPN | CA | RepVGG | mAP | Params/$10^6$ | GFLOPs/$10^9$ |
|---|---|---|---|---|---|---|
| YOLOv5 | | | | 0.787 | 7.01 | 15.8 |
| Improvements 1 | √ | | | 0.804 | 7.16 | 16.4 |
| Improvements 2 | √ | √ | | 0.813 | 7.48 | 17.9 |
| Improvements 3 | √ | √ | √ | 0.817 | 7.65 | 18.2 |

### 4.5 Comparative Experiments and Analyses

In order to further prove the superiority of the improved algorithm, the improved YOLOv5 algorithm is used to do comparison experiments with the original YOLOv5 algorithm, SSD, YOLOv7, and YOLOv8, all the algorithms are trained and validated on the same dataset, and the experimental environments and equipments are the same, and the comparison is done with the same evaluation indexes. The experimental results are shown in Table 2, the SSD algorithm with lower mAP has poorer recognition ability for dense targets, YOLOv8 accuracy is lower than YOLOv5 and YOLOv7 accuracy is 0.3% higher than YOLOv5, but the number of parameters and the amount of computation is larger than that of YOLOv5 and it is difficult to meet the real-time requirements. The mAP of the improved model is significantly higher than that of other mainstream algorithms, and the mAP is up to 0.817, which is the optimal model.

**Table 2.** Comparative experimental results

| Model | mAP | Params/$10^6$ | GFLOPs/$10^9$ |
|---|---|---|---|
| Improve model | 0.817 | 7.65 | 18.2 |
| SSD | 0.31 | 23.87 | 30.47 |
| YOLOv5 | 0.787 | 7.01 | 15.8 |
| YOLOv7 | 0.790 | 36.48 | 103.2 |
| YOLOv8 | 0.773 | 3.01 | 8.2 |

## 5. Conclusion

Recognition of morning glory peppers in densely occluded environments has the problem of causing leakage and low accuracy, an improved YOLOv5s recognition algorithm is proposed. The convolutional kernel is replaced with the improved RepVGG module; the CA attention module is used to fuse with the Bottleneck module in the C3 module, so that the model focuses on the focus area to improve the accuracy; and the weighted multi-scale fusion using BiFPN enhances the feature fusion of the occluded target. After experimental verification, the improved algorithm can quickly and accurately identify the occluded chilli and improve the accuracy, improve the leakage detection problem and enhance the robustness of the model.

## References

[1] Balejian Madiniyanti, Bwaibuyong Abra. Analysis of comparative advantages and influencing factors of China's fruit export trade[J]. World Agriculture,2019(7):57-68.

[2] Zhong Shihao. Research on target recognition and localisation algorithm of cluster Chaotian pepper based on deep learning [D]. Guizhou:Guizhou Normal University,2023.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202405_10(5).0038

[3] Hu Shuangji, Chen Yongcheng,Yuan Yinxia, et al. Research status and prospect of chilli harvester[J]. Agricultural Mechanisation Research,2011,33(8):237-240.

[4] Gou Yuanmin, Yan Jianwei, Zhang Fugui, et al. Research progress of vision system and manipulator for fruit picking robot[J]. Computer Engineering and Application,2023,59(9): 13-26.

[5] Yu Siqian, Zhao Qirong, Lin Jiachen, et al. Detection of walnut shell defects based on deep learning[J]. Journal of Jilin Institute of Chemical Technology,2022,39(9):80-85.

[6] REN S Q, HE K M, GIRSHICK ROSS, et al. Faster R-CNN:Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017, 39(6):1137-1149.

[7] Wu Chenxi,Ying Baosheng,Xu Xiaowei,et al. Road small target detection algorithm based on improved single-step multi-frame target detection[J]. Science Technology and Engineering,2023,23(5):2051-2058.

[8] Luo Huilan,Chen Hongkun. A review of deep learning based target detection research[J]. Journal of Electronics,2020,48(6):1230-1239.

[9] Shang Fengnan, Zhou Xuecheng, Liang Yingkai, et al. An improved YOLOX-based method for dragon fruit detection in natural environment[J]. Intelligent Agriculture (in Chinese and English),2022,4(3):120-131.

[10] Lu Wenkang, Lu Shenglian, Liu Binghao, et al. Research on orchard citrus detection method based on improved YOLOv4[J]. Journal of Guangxi Normal University (Natural Science Edition),2021,39(5):134-146.

[11] Liu Qi, Yin Gang, Wang Ying, et al. Design of deep learning-based algorithm for water surface floating object recognition[J]. Journal of Jilin Institute of Chemical Technology,2022,39(7):28-33.