Standard Issue – Article

Journal of Agricultural Engineering

https://www.agroengineering.org/

Fast identification of tomatoes in natural environments by improved YOLOv5s

Hongbo Wang, Zhicheng Xie, Yongzheng Yang, Junmao Li, Zilu Huang, Zhihong Yu

Please cite this article as doi: 10.4081/jae.2024.1588

Submitted: 18/11/2023
Accepted: 06/06/2024

Fast identification of tomatoes in natural environments by improved YOLOv5s

Hongbo Wang, Zhicheng Xie, Yongzheng Yang, Junmao Li, Zilu Huang, Zhihong Yu

Inner Mongolia Agricultural University Ringgold Standard Institution, Hohhot, China

Correspondence: Hongbo Wang, Inner Mongolia Agricultural University - College of Mechanical and Electrical Engineering, No. 306 Zhaowuda Rd Hohhot 010018, China.
E-mail: wanghb@imau.edu.cn

Key words: tomatoes; YOLOv5s; fast identification.

Acknowledgments: the authors are grateful for the support of the National Natural Science Foundation of China, Grant No. 52265035; Inner Mongolia Science and Technology Innovation Guidance Reward Fund of China, Grant No. Kcjl-202205.

Conflict of interest: the authors declare no potential conflict of interest.

Abstract

Real time recognition and detection of tomato fruit maturity is a key function of tomato picking robots. Existing recognition and detection algorithms have slow speed and low recognition accuracy for small tomatoes. Here, a tomato fruit maturity detection model YOLOv5s3 based on improved YOLOv5s was proposed and its accuracy was verified through comparative experiments. On the basis of YOLOv5s, an SC module was proposed based on channel shuffle packet convolution. Then, A C3S module is constructed, which replaced the original C3 module with this C3S module to reduce the number of parameters while maintaining the feature expression ability of the original network. And a 3-feature fusion FF module was put forward, which accepted inputs from three feature layers. The FF module fused two feature maps from the backbone network. The C2 layer of the backbone was integrated, and the large target detection head was removed to use dual head detection to enhance the detection ability of small targets. The experimental results showed that the improved model has a detection accuracy of 94.8%, a recall rate of 96%, a parameter quantity of 3.02M, and an average accuracy (mAP0.5) of 93.3% for an intersection over union (IoU) of 0.5. The detection speed reaches 9.4ms. It can quickly and accurately identify the maturity of tomato fruits, and the detection speed is 22.95%, 33.33%, 48.91%, 68.35%, 15%, and 25.98% higher than the original YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x, YOLOv5n, and YOLOv4, respectively. The real-time testing visualization results of different models indicated that the improved model can effectively improve detection speed and solve the problem of low recognition rate for small tomatoes, which can provide reference for the development of picking robots.

Introduction

With the rapid development of science and technology, the level of agricultural informatization and intelligence has been greatly improved. Currently, visual detection technology for fruits and vegetables, which relies on deep learning, has high recognition accuracy and rich feature extraction (Lü et al., 2019). In recent years, more and more scholars have combined deep learning with agriculture to apply it to fields such as picking and harvesting (Cámara-Zapata et al., 2019; Ma et al., 2021; Xu et al., 2022).

Due to the inconsistent maturation time of most fruits and vegetables during their growth process and the characteristic of batch maturation, the judgment of fruit maturity is an important link in automated picking (Liu et al., 2022). Planting tomatoes in greenhouses can reduce the impact of pests and diseases, but the cost of manual picking is very high. In terms of smart agriculture, tomatoes are one of the important crops in China and an important economic crop, ranking first in terms of planting area and yield in the world. As far as Inner Mongolia Autonomous Region is concerned, the yield

per acre can reach 20000 kilograms, especially tomatoes from Wuhai region, which are economically advantageous crops in Inner Mongolia Autonomous Region. Therefore, people's demand for high-quality tomatoes will increase with our living standards. Therefore, in order to achieve automatic tomato picking, it is crucial to conduct precise identification research on tomato fruits (Yang et al., 2022).

In terms of automated tomato fruit picking, the most important point is the precise detection of tomato fruit maturity, providing favorable guarantees for subsequent automatic picking and path planning. And with the emergence of convolutional neural networks, deep learning technology and its excellent feature extraction and generalization capabilities have gradually been applied to the recognition of fruits, vegetables, and fruits (Shang et al., 2022). The object detection algorithms for deep learning can be divided into two types: one stage and two stage. Two stage refers to the candidate boxes generated by the algorithm for a series of samples, which are then classified using convolutional neural networks. The detection accuracy is high but the real-time performance is poor, represented by R-CNN, Faste R-CNN, Faster R-CNN, etc.; One stage does not need to generate candidate boxes, but directly transforms the problem of locating the target border into a regression processing problem, which has a fast detection speed and represents the YOLO series of algorithms.

In the field of fruit and vegetable recognition and detection, many experts and scholars have conducted relevant research on two types of object detection algorithms. In the study of target detection algorithms for two-stage, Chen et al. (2021) proposed an improved Faster R-CNN model to recognize cotton top buds in field environments. The guided anchoring and GROIE mechanisms were fused to enhance the model's recognition ability for cotton top buds, with a recognition accuracy of 98.1% and a processing frame rate of 10.3 frames/s. Yao et al. (2020) proposed a feature extraction network based on ResNetXt101 for RetinaNet to identify rice canopy pests. By normalizing and improving the feature pyramid structure, the recognition accuracy reached 93.76%. Chen et al. (2022) embedded Gabor into Faster R-CNN and proposed a two-stage training method based on genetic algorithm and backpropagation to train a new Faster GG-R-CNN model, with an average accuracy of 94.57%. Zhang et al. (2021) proposed an improved Faster R-CNN based rice ear detection method to monitor the number of rice ears in the monitoring area. To address the problem of small rice ear targets, they applied Inception_ Introducing hollow convolution for optimization based on ResNet v2; For the problem of significant differences in rice panicles at different growth stages, K-means clustering was designed at the label box scale to provide prior knowledge for candidate region generation networks, thereby improving detection accuracy. The mAP reached 80.3%. Seo et al. (2021) designed and developed a real-time robot detection system based on Faster R-CNN and used color tone values to develop image-based maturity standards for tomato fruits. The recognition

accuracy reached 90.2%. This type of two-stage object detection algorithm based on region extraction has high detection accuracy, but it is difficult to achieve real-time detection and cannot meet the real-time and efficient requirements of agricultural intelligent equipment.

In the study of one stage object detection algorithms, Yang et al. (2022) proposed a BCo YOLOv5 network model to identify and detect fruit targets in orchards. YOLOv5s was used as the basic model for feature image extraction and object detection, and BCAM (Bidirectional Cross Attention Mechanism) was introduced. BCAM was added between the backbone network and bottleneck network of the YOLOv5s basic model, mapping the BCo YOLOv5 network model to 97.70%. Two experiments were designed based on data from aquaculture ponds to test the performance of the proposed model. The average accuracy and recall of DCM-ATM-YOLOv5 are 97.53% and 98.09%, respectively. Sun et al. (2022) proposed an improved YOLOv5s model for identifying apple diseases. By adding a phantom structure and adjusting the overall width of the feature map, a small baseline model was obtained, and TR2 was used as the detection head to improve the model's ability to obtain information. The improved model had a recognition accuracy of 90.9% and a recognition speed of 0.065/s per image. This type of two-stage object detection algorithm based on regression maintains high detection accuracy while also having high detection speed, which can achieve real-time detection and meet the needs of agricultural intelligent equipment.

At present, for the maturity detection of fruits and vegetables in deep learning, tomato fruits generally grow densely in natural environments, with overlapping or obstructing each other between fruits and leaves; The growth environment where tomatoes are grown is open and unshaded, and under the greenhouse cultivation mode, the fruit is prone to leaf obstruction and uneven light, as well as the recognition of fruit characteristics during the color transition period. These factors pose certain difficulties for accurate recognition of tomato fruit maturity.

Therefore, for the recognition and detection of tomato fruit maturity in complex natural environments, this study proposes an improved YOLOv5s tomato fruit maturity detection method.

By improving the backbone network, neck, and prediction layer structure of YOLOv5s, combined with certain data processing methods, rapid and accurate recognition and detection of tomato fruit in natural environments can be achieved. This study can provide theoretical basis and technical support for the tomato fruit picking robot and system.

Materials and Methods

Data collection and dataset construction

The determination and recognition of fruit maturity is a necessary condition for mechanized harvesting. The maturity of fruits and vegetables can affect their use and harvesting time, so selective

harvesting is necessary.

The current national standard GH/T1193-2021 divides the maturity of tomatoes into immature stage, green stage, discoloration stage, early red stage, middle red stage, and late red stage based on their color and size. The fruit in the middle and late red stages has a red surface of 40% to 60% and 70% to 100%, respectively. In a greenhouse, each tomato tree usually has at least three different maturities (green maturity, color transition, and red maturity), and only tomatoes in the middle and late stages of red maturity are picked. 1) Tomatoes during the green maturity period are used for long-term storage or long-distance transportation due to their solid fruit and strong disease resistance; 2) During the color transition period, tomatoes are harvested for short-term storage or close transportation; 3) Due to its perishability during the maturity period, fruits that are about to go on sale are harvested (Xiao et al., 2015). The tomatoes in the middle and late stages of red maturity are referred to as mature tomatoes, while other mature tomatoes are collectively referred to as immature tomatoes. Therefore, in order to pick the most suitable tomatoes for transportation and sales, it is very important to accurately determine the maturity of tomatoes.

The tomato image data used in this study was collected from the picking garden of Tumote Left Banner Farm in Hohhot City, using standardized planting methods. In order to enhance the robustness, generalization, and applicability of the model, all images were taken between 0.5-1m from the tomato, and 2500 images with a resolution of were selected to create a dataset, as shown in Figure 1.

To ensure the accuracy of image data parameters, the labeling was used to manually annotate the image data before training the model. Pascal voc annotation format was used. In this study, the annotation target, namely the tomato fruit to be detected, was divided into two categories, with ripe tomatoes labeled as Ripe and immature tomatoes labeled as Immature. Use the smallest bounding rectangle on all four sides of the tomato as a real box to reduce unnecessary pixels in the background. Once annotated, an. xml file will be generated. And Gaussian blur, random brightness, random contrast, random tone, and random saturation data augmentation were applied to the data to double the original data. The data was divided into a training set and a testing set in an 8:2 ratio, with 30% of the testing set used as the validation set for cross validation of model training. The final training set consisted of 2000 images, 500 images, and 150 images.

Construction of tomato maturity detection model based on improved YOLOv5s

The YOLOv5 algorithm is an improvement on the YOLOv4 algorithm. YOLOv5 has accurate, real-time, and efficient detection results. Compared with YOLOv4, it achieves lightweight and improved detection accuracy of the model while ensuring accuracy. YOLOv5 has 5 different

structures for different conditions, namely YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x, and YOLOv5n. The several models only differ in depth and width, The basic structure is composed of four parts: input, backbone network, neck, and head. In order to meet the demand for lightweight, the YOLOv5s detection model with the lowest complexity is selected, which can achieve faster recognition speed and smaller storage occupation while ensuring high detection accuracy, which is conducive to the construction of mobile devices. (Yu et al., 2022; Jiang et al., 2022; Wang et al., 2022).

The main structure of YOLOv5s consists of four parts: input, backbone, neck, and head (Qi et al., 2022). In the model training stage, the input side proposes Mosaic data enhancement, adaptive anchor box calculation, and adaptive image scaling, optimizing the image processing strategy and anchor box generation mechanism. The backbone network is used to extract information from images, and using the CSPParknet53 structure can reduce the number of network parameters and effectively solve the problem of gradient vanishing (Yu et al., 2021; Qiao et al., 2022). The neck part uses the network structure of feature pyramid and path aggregation structure to enhance features, thereby enriching the extracted feature network information. The prediction part obtains network output, and the head uses the previously extracted features to make predictions.

The target detection of tomato maturity in natural environments has been studied by us. Based on YOLOv5s, an improved YOLOv5s3 network model structure is proposed. Due to the dense growth of tomato fruits and the overlap and occlusion between fruits and leaves, there may be situations where the feature representation is not obvious, and tomato fruits are generally small targets. Therefore, it is necessary to improve the network structure, Improve the ability of feature extraction, significantly increase detection speed, reduce parameter count, and achieve lightweight while ensuring high accuracy. Based on the above considerations, the main improvement points of the model include: introducing C3S module to replace C3 module in the backbone network, reducing the number of parameters and enhancing the ability of feature extraction; In the neck network and prediction layer, a three-feature fusion FF module is proposed, which accepts inputs from three feature layers and removes the detection head of large targets to further improve detection speed. The improved YOLOv5s3 model is expected to further reduce the number of parameters and improve detection speed while maintaining high accuracy. The improved structure of YOLOv5s3 is shown in the figure 2

C3S module

The original C3 module has a large amount of 3x3 convolution parameters, so the C3S (Figure 3) module is proposed. The C3S module is composed of two 1x1 convolutions and two SC modules.

The SC module (Figure 4) proposes a packet convolution based on Channel Shuffle to reduce the number of parameters while maintaining the feature expression ability of the original network. The channel exchange is interspersed between two group convolutions GConv to achieve information exchange between different groups of channels. Based on the SC module, a C3S module is constructed, replacing the C3 module in the network.

The operation process of C3S module is as follows:

1) The input initial feature information undergoes a 1x1 convolution for channel split, and the information is distributed to two branches;

2) Reduce the number of input feature information channels by half, and perform feature extraction on the feature map with half of the channels using two SC modules;

3) Finally, perform 1x1 convolution dimensionality reduction on all outputs, so that the feature information size of the output C3S module is the same as that of the input C3S module.

Compared to the original YOLOv5s model, the introduction of C3S module enhances the ability of network learning, reduces the computational parameters and memory size in a lightweight way, improves the detection speed of the model, and enhances the ability to extract tomato details in the image. By learning more accurate tomato features, effective detection of overlapping and blocked tomato fruits in the image is achieved.

Feature fusion dual head detection

The tomato fruits grown in greenhouse cultivation mode are open and unobstructed, and the image background contains complex backgrounds such as direct light, shadows, and branches and leaves, which affects the accuracy of the model's fruit detection. In order to further enhance the model's ability to extract fruit features from complex backgrounds, this paper introduces a 3-feature fusion FF module in the neck network of YOLOv5s to improve the model and further adapt to the rapid detection of tomato fruits in natural environments.

Due to the limitations of the Concat, which accepts inputs from only two feature maps, this paper proposes a 3-feature fusion FF module that accepts inputs from three feature layers. The FF module aligns different feature map sizes through upsampling Upsample and MaxPool, and adaptively fuses multiple input features through Concat concatenation and 1x1 convolution. The FF module is different from the original Concat, The Concat only integrates a single feature map from the backbone network, while the FF module integrates two feature maps from the backbone network and enhances the small target detection ability by integrating the C2 layer of the backbone. Additionally, there are many small targets in the dataset, but no large targets, so the detection head for large targets is removed. The FF module structure is shown in Figure 5.

Compared to the original YOLOv5s model, by introducing the FF module, the adaptive fusion of three feature maps was achieved, further enhancing the model's ability to detect small targets, enhancing the model's ability to extract fruit features in complex backgrounds, improving the robustness of the model, improving the accuracy of the model for fruit detection, reducing the number of feature parameters, and further improving the detection speed.

Experimental environment

Vscode was used by us to construct and improve the YOLOv5s network model. The operating system used for testing and training is Windows 10, the processor is 11th Gen Intel (R) Core (TM) i7-11800H @ 2.30GHz, 2.30 GHz, the memory is 32GB, the graphics card is NVIDIA GeForce RTX 3060 Laptop GPU, the graphics memory is 16GB, the programming language is Python 3.8.2, and the training is conducted using the Python deep learning framework.

Evaluation indicators

Five indicators were used to evaluate the tomato fruit maturity detection model, namely accuracy (P), recall rate (R), average accuracy (mAP), number of model parameters, and detection speed. The intersection over union (IoU) is used for accuracy evaluation. The calculation formulas for P, R, mAP, and IoU are as follows (Redmon et al., 2016; Jiang et al., 2018).

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$IoU = \frac{S_{\cap}}{S_{\cup}} \tag{3}$$

$$AP = \int_0^1 P(R)dR \tag{4}$$

$$mAP = \frac{\sum_{i=1}^{Nc} AP_i}{N_c} \tag{5}$$

In equation (1), TP represents the number of positive samples predicted to be positive, FP represents the number of negative samples predicted to be negative, FN represents the number of positive samples predicted to be negative, IoU is the ratio of the intersection and union of predicted border and real border, AP represents the area covered by the P(R) curve made with Recall value as the X-axis and Precision value as the Y-axis. It measures the recognition accuracy of a certain category. In equation (5), mAP represents the average value of each category's AP, and Nc represents the number of categories, which measures the average level of good or bad across all categories.

Model training

For the improved YOLOv5s3 tomato fruit maturity detection model, set the initial learning rate of 1e-2, input image size, and train a total of 300 rounds using the Random Gradient Descent (SGD) network training.

During the training process, the loss function, accuracy (P), recall (R), and P-R curve of the model are recorded as shown in Figure 6. From Figure 6, it can be seen that the loss function decreases rapidly in the early stages of training, and the overall fluctuation is small. At this time, the learning efficiency of the model is high, and the loss value converges quickly; When the iteration reaches around 250 times, the loss value converges to around 0.025; After 300 iterations, the P-value of the model is 94.8%, R-value is 96%, and mAP0.5 is 93.3%.

Results

Comparison of ablation experimental performance

In order to better analyze the impact of improving the YOLOv5s model on the detection of tomato fruit maturity, YOLOv5s was used as the benchmark model to evaluate the optimization effect of each improvement point through accuracy, recall, mean accuracy (mAP0.5), detection time, and floating-point computation. The structure of the ablation experiment is shown in the table, with bold font being the optimal value.

Based on the analysis of the table below, it can be seen that the optimized model 1 achieved an average detection accuracy and speed of 90.3% and 98 fps respectively by modifying the backbone network of the YOLOv5s model, which increased by 0.4% and 16 fps compared to the original model; Optimizing Model 2 by modifying the neck and detection layer of the YOLOv5s model and using dual head detection, the average accuracy and speed of detection were improved by 0.8% and 11 fps, respectively; The final accuracy and speed of the YOLOv5s3 model were 93.3% and 106 fps, respectively. The recall rate increased by 1.9%, the average accuracy increased by 0.5%, and the detection speed increased by 22 fps.

The results show that introducing the C3S module to the original model for parameter reduction can significantly improve the detection speed of the improved model, but the accuracy will be slightly reduced. Furthermore, introducing the three-feature fusion dual head detection and FF module further improves the accuracy of the improved model, and the detection speed improvement is relatively slow.

Performance comparison of different target detection models

In order to evaluate the detection effect of the proposed YOLOv5s3 model on tomato fruit maturity, the original YOLOv5s, YOLOv5m, YOLOV5l, YOLOv5x, YOLOV 5n, and YOLOv4 (Zhou et al., 2022; Wang et al., 2021) were selected for performance comparison. Using 1.5 section images, the original YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x, YOLOv5n, and YOLOv4 were trained for 300 training rounds. We will only use the test set to conduct performance evaluation tests on the six trained detection models mentioned above. The performance comparison of the six detection models is shown in the table below.

By analyzing the test results in the table above, it can be concluded that the accuracy of the proposed improved model YOLOv5s3 in detecting tomato fruit maturity is 1 percentage point lower than YOLOv5x and 0.8 percentage points higher than YOLOv5s.The model parameter quantity is 56%, 85.73%, 93.37%, 96.51%, 25.98%, and 96.81% lower than the original YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x, YOLOv5n, and YOLOv4, respectively. The detection speed is 56%, 85.73%, 93.37%, 96.51%, 25.98%, and 96.81% lower than the original YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x, and YOLOv5x, respectively YOLOv5n and YOLOv4 are 22.95%, 33.33%, 48.91%, 68.35%, 15%, and 25.98% faster.

The reasons for the above results are as follows:

1) The C3S module based on YOLOv5s and 3 feature fusion dual head detection were introduced into the model, reducing the complexity of the model and significantly reducing the number of parameters. Compared with YOLOv5s, the model size was significantly reduced. In addition, the detection accuracy, recall rate, and accuracy have slightly improved, and the detection speed has significantly improved.

2) The model in this article is based on YOLOv5s (Du et al., 2022). Compared to YOLOv4 (Peng et al., 2022), YOLOv5s incorporates Backbone and Neck with an improved CSP structure, and the anchor box is automatically learned based on the training dataset. However, YOLOv4 does not have an adaptive anchor box, which significantly improves the recall, accuracy, and detection speed without significantly reducing the detection accuracy, and significantly reduces the parameter size of the model.

The results indicate that our proposed improved model YOLOv5s3 outperforms other object detection models in terms of overall performance. It can achieve accurate and rapid detection of tomato fruit maturity, and achieve lightweight of the model, which is conducive to migration and application. It can provide certain technical support for the development of tomato fruit picking equipment.

Evaluation of detection effect

The recognition results of different models are shown in Figure 7. In order to quickly analyze the recognition situation, mature tomatoes are circled in a pink bounding box with a "Ripe" label, and immature tomatoes are circled in a red bounding box with an "Immature" label. In order to evaluate the effectiveness of our improved model YOLOv5s3 in detecting tomato fruit maturity, the detection accuracy of the training model was verified through 500 test set images. The results showed that the P-value, R-value, mAP0.5 value, and model parameter quantity of YOLOv5s3 for detecting tomato fruit maturity were 94.8%, 96%, 93.3%, and 3.02M, respectively. The detection speed was 9.4ms.

The detection example is shown in Figure 7, where each mature and immature tomato fruit can be accurately identified. The average confidence level of mature tomato fruit detection is above 0.95, while the average confidence level of immature tomato fruit detection is above 0.91, indicating that the model can effectively detect the maturity of tomato fruit and has a high confidence level. In addition, for images collected in more complex scenes, such as overlapping images, leaf occluded images, and densely distributed fruits, the improved model YOLOv5s3 in this paper can achieve effective detection.

Analysis of misdetection and misdetection

When using the YOLOv5s3 model in this study to detect tomato fruits in the validation set images, there may still be missed and false detections, as shown in Figure 8. From the figure, it can be seen that some tomato fruit maturity detection results have missed and false detections, and there are also cases where leaves are mistakenly considered as immature tomato fruits and tomatoes during the color transition period are considered mature tomato fruits.

From Figure 8, it can be seen that some mature tomato fruits were missed detection (as shown in the yellow circles in Figure 8a and 8b), while there were also cases where the leaves were mistakenly detected as immature tomato fruits (as shown in the yellow circles in Figure 8c) and tomato fruits during the color transition period were mistakenly detected as mature tomato fruits (as shown in the yellow circles in Figure 8d). The possible reasons for the missed detection and missed detection are as follows:

1) The YOLOv5s3 model has good detection performance for sparsely distributed fruits, but under natural conditions, the distribution of fruits is generally relatively dense, and there may be occlusion between fruits. As shown in 8a and 8b, when the occlusion is severe, the tomato fruit in the upper layer will cover most of the features of the tomato fruit in the lower layer. This will result in relatively obvious features of the upper layer tomato fruit, making it difficult for the model to accurately identify the features of the lower layer tomato fruit. When extracting the features of the lower layer tomato fruit, it is often difficult. This will cause the model to only successfully detect the maturity of

the upper layer tomato fruit when detecting it, resulting in the inability to detect the lower layer tomato fruit and the occurrence of missed detection.

2) The image quality of tomato fruit images collected in natural environments is influenced by various factors such as light intensity and weather. Uneven distribution of light intensity and leaf occlusion can lead to low contrast between immature tomato fruits obstructed by leaves and the leaf background, making it difficult for the model to extract features of immature tomato fruits. When there is obstruction between leaves and fruits, The upper leaves will also cover most of the features of the lower fruit, causing the immature tomato fruit and leaves to exhibit similar features. During model detection, the lower immature tomato will be detected as the background or the leaves will be detected as immature tomato fruit, as shown in 8c, resulting in false detection.

3) During the ripening process, tomato fruits undergo a color transition period, with the color changing from green to red. Red tomatoes are defined as mature tomato fruits, while green tomatoes are defined as immature fruits. Therefore, during the color transition period, the fruit exhibits similar characteristics to mature tomatoes, as shown in Figure 8d. However, at this time, the tomato fruit is not fully mature and does not meet the harvesting conditions, Therefore, this study defines the color changing tomato fruit as an immature tomato fruit, in order to meet the picking standards and not accidentally pick the color changing tomato fruit. However, due to the similar characteristics between fruits during the color transition period and mature fruits in such cases, the model is prone to mistakenly defining such fruits as mature fruits when extracting the characteristics of fruits during the color transition period, as shown in 8d, resulting in false detection.

Conclusions

The weight file of the current neural network model is too large, and the real-time detection speed is relatively slow. On the basis of our improved YOLOv5s model, we have constructed an improved tomato fruit maturity detection model YOLOv5s3, which achieves rapid and accurate detection of tomato fruit maturity in natural environments. The main conclusions are as follows:

1) In this paper, an improved YOLOv5s3 model based on the Yolov5s model is proposed. The average accuracy (P) of the model for detecting tomato fruit maturity is 94.8%, and the recall rate (R) is 96%. The introduction of C3S module in Backbone replaces the traditional C3 module to reduce the number of parameters while maintaining the feature expression ability of the original network. The FF module of three feature fusion is introduced in Neck and prediction, and the dual-head detection is adopted to enhance the extraction ability of features and strengthen the detection ability of small targets. The mean average precision with IoU of 0.5 (mAP0.5) reached 93.3% when the intersection to union ratio was 0.5. The model had a parameter size of 3.02M and a detection

speed of 9.4ms. Through the improvement of the model, a fast and accurate detection of tomato fruit maturity in natural environments was achieved.

2) The performance of YOLOv5s3 proposed by us was compared with four other improved models based on YOLOv5, the original YOLOv5s model, and YOLOv4 model. The model parameters are 56%, 85.73%, 93.37%, 96.51%, 25.98%, and 96.81% lower than the original YOLOv5s, YOLOv5m, YOLOv4, YOLOv5x, YOLOV5 n, and YOLOv4.The detection speed is 22.95%, 33.33% faster than the original YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x, YOLOv5n, and YOLOv4, respectively. 48.91%, 68.35% 15%, 25.98%, the results showed that the YOLOv5s3 model ouTPerformed the other models in terms of detection speed and parameter quantity, while the detection accuracy difference was not significant. This indicates that the improved model in this paper has a good detection effect on tomato fruit maturity, with a detection accuracy of mAP0.5 value of 93.3%.

The YOLOv5s3 model proposed by us has some omissions and errors in detecting tomato fruits that grow too densely, have severe occlusion between fruits, and between leaves and fruits. Subsequent work will continue to seek new improvement plans to further optimize the detection effect of tomato fruit maturity in complex situations.

References

Cámara-Zapata J.M., Brotons-Martínez J.M., Simón-Grao S., Martinez-Nicolás J.J., García-Sánchez F. 2019. Cost–benefit analysis of tomato in soilless culture systems with saline water under greenhouse conditions. J. Sci. Food Agric. 99:5842-5851.

Chen K.Y., Zhu L.F., Song P., Tian X.M., Huang C.L., Nie X.H., Xiao A.L., He L.R. 2021. Recognition of cotton terminal bud in field using improved Faster R-CNN by integrating dynamic mechanism. Transactions of the Chinese Society of Agricultural Engineering. 37(16):161-168.

Chen M.Q, Yu L.J., Zhi C., Sun R.J., Zhu S.W., Gao Z.Y., Ke Z.X., Zhu M.Q., Zhang Y.M. 2022. Improved faster R-CNN for fabric defect detection based on Gabor filter with Genetic Algorithm optimization. Computers in Industry. 134:103551.

Du F.J., Jiao S.J. 2022. Improvement of lightweight convolutional neural network model based on YOLO algorithm and its research in pavement defect detection. Sensors. 22:3537.

Jiang B.R., Luo R.X., Mao J.Y., Xiao T.T., Jiang Y.N. 2018. Acquisition of localization confidence for accurate object detection. European Conference on Computer Vision (ECCV) 2018. 1807.11590.

Jiang T., Li C., Yang M., Wang Z. 2022. An improved YOLOv5s algorithm for object detection with an attention mechanism. Electronics. 11:2494.

Liu X.G., Fan C., Li J.N., Gao Y.L., Zhang Y.Y., Yang Q.L. 2020. Identification method of strawberry

based on convolutional neural network. Transactions of the Chinese Society for Agricultural Machinery. 51(2):237-244.

Lü S.L., Lu S.H., Li Z., Hong T.S., Xue Y.J., Wu B.L. 2019. Orange recognition method using improved YOLOv3-LITE lightweight neural network. Transactions of the Chinese Society of Agricultural Engineering. 35(17):205-214.

Ma L., Guo X.L., Zhao S.K., Yin D.D., Fu Y.Y., Duan, P.Q., Wang B.B., Zhang L. 2021. Algorithm of strawberry disease recognition based on deep convolutional neural network. Complexity. 2021:6683255.

Qi J.T., Liu X.N., Liu K., Xu F.R., Guo H., Tian X.L., Li M., Bao Z.Y., Li Y. 2022. An improved YOLOv5 model based on visual attention mechanism: Application to recognition of tomato virus disease. Computers and Electronics in Agriculture. 194:106780.

Qiao Y., Hu Y., Zheng Z., Yang H., Zhang K., Hou J., Guo J. 2022. A counting method of red jujube based on improved YOLOv5s. Agriculture. 12:2071.

Redmon J., Divvala S., Girshick R., Farhadi A. 2016. You only look once: unified, real-time object detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition. 2016:779-788.

Seo D., Cho B-H., Kim K-C. 2021. Development of monitoring robot system for tomato fruits in hydroponic greenhouses. Agronomy. 11:2211.

Shang Y.Y., Zhang Q.R., Song H.B. 2022. Application of deep learning using YOLOv5s to apple flower detection in natural scenes. Transactions of the Chinese Society of Agricultural Engineering. 38(9): 222-229.

Sun F.G., Wang Y.L., Lan P., Zhang X.D., Chen X.D., Wang Z.J. 2022. Identification of apple fruit diseases using improved YOLOv5s and transfer learning. Transactions of the Chinese Society of Agricultural Engineering. 38(11):171-179.

Wang F., Sun Z., Chen Y., Zheng H., Jiang J. 2022. Xiaomila green pepper target detection method under complex environment based on improved YOLOv5s. Agronomy. 12:1477.

Wang L.S., Qin M.X., Lei J.Y., Wang X.F., Tan K.Z. 2021. Blueberry maturity recognition method based on improved YOLOv4-Tiny. Transactions of the Chinese Society of Agricultural Engineering 37(18):170-178.

Xiao Q.M., Niu W.D., Zhang H. 2015. Predicting fruit maturity stage dynamically based on fuzzy recognition and color feature. Proceedings of 2015 IEEE 6th International Conference on Software Engineering and Service Science. 2015:968-972.

Xu W.C., Yan Z. 2022. Research on strawberry disease diagnosis based on improved residual network recognition model. Mathematical Problems in Engineering. 2022:6431942.

Yang J., Qian Z., Zhang Y.J., Qin Y., Miao H. 2022. Real-time recognition of tomatoes in complex

environments based on improved YOLOv4-tiny. Transactions of the Chinese Society of Agricultural Engineering. 38(9):215-221.

Yang R.L., Hu Y.W., Yao Y., Gao M., Liu R.M. 2022. Fruit target detection based on BCo-YOLOv5 model. Mobile Information Systems. 2022:8457173

Yao Q., Gu J.L., Lv J., Guo L.J., Tang J., Yang B.J, Xu W.G. 2020. Automatic detection model for pest damage symptoms on rice canopy based on improved RetinaNet. Transactions of the Chinese Society of Agricultural Engineering. 36(15):182-188.

Yu X.H., Kong D.Y., Xie X.X., Wang Q., Bai X.W. 2022. Deep learning-based target recognition and detection for tomato pollination robots. Transactions of the Chinese Society of Agricultural Engineering. 38(24):129-137.

Yu Y., Zhao J., Gong Q., Huang C., Zheng G., Ma J. 2021. Real-time underwater maritime object detection in side-scan sonar images based on transformer-YOLOv5. Remote Sens. 13:3555.

Zhang F., Chen Z.J., Bao R.F., Zhang C.C., Wang Z.H. 2021. Recognition of dense cherry tomatoes based on improved YOLOv4-LITE lightweight neural network. Transactions of the Chinese Society of Agricultural Engineering. 37(16):270-27

Zhang Y.Q., Xiao D.Q., Chen H.K., Liu Y.F. 2021. Rice panicle detection method based on improved faster R-CNN. Transactions of the Chinese Society for Agricultural Machinery. 52(8):231-240.

Zhou G.H., Ma S., Liang F.F. 2022. Recognition of the apple in panoramic images based on improved YOLOv4 model. Transactions of the Chinese Society of Agricultural Engineering. 38(21):159-168.

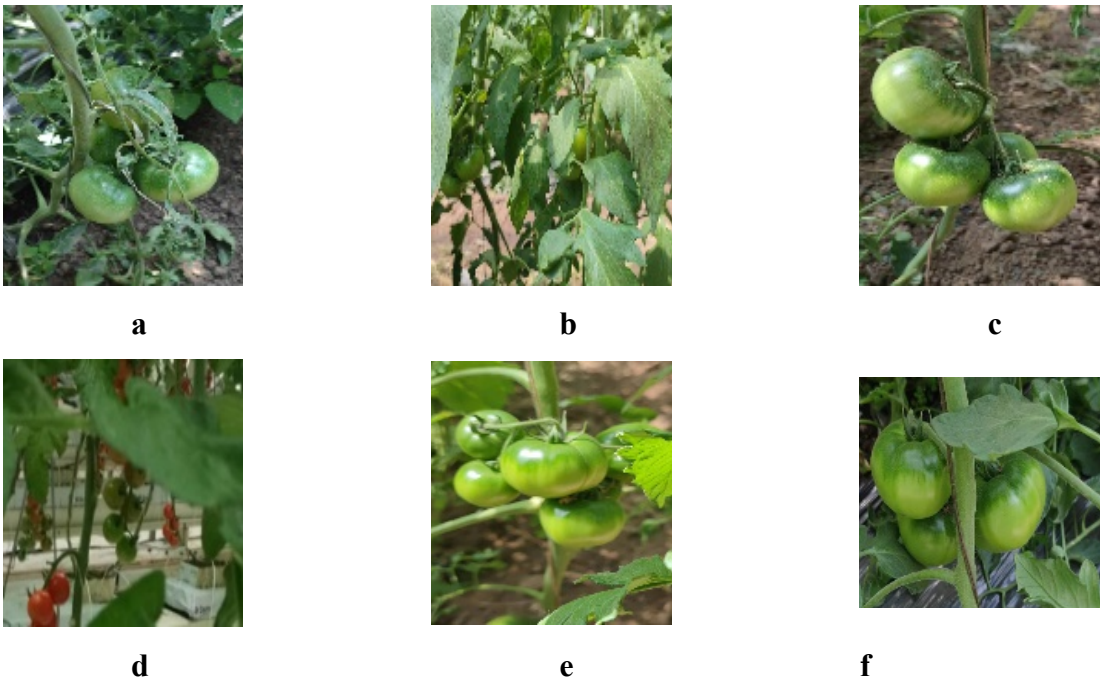


Figure 1. Tomato dataset samples under different scenarios. a) No occlusion; b) Occlusion; c) Clusters of tomatoes; d) Multiple targe; e) Adequate lighting; f) Insufficient light.

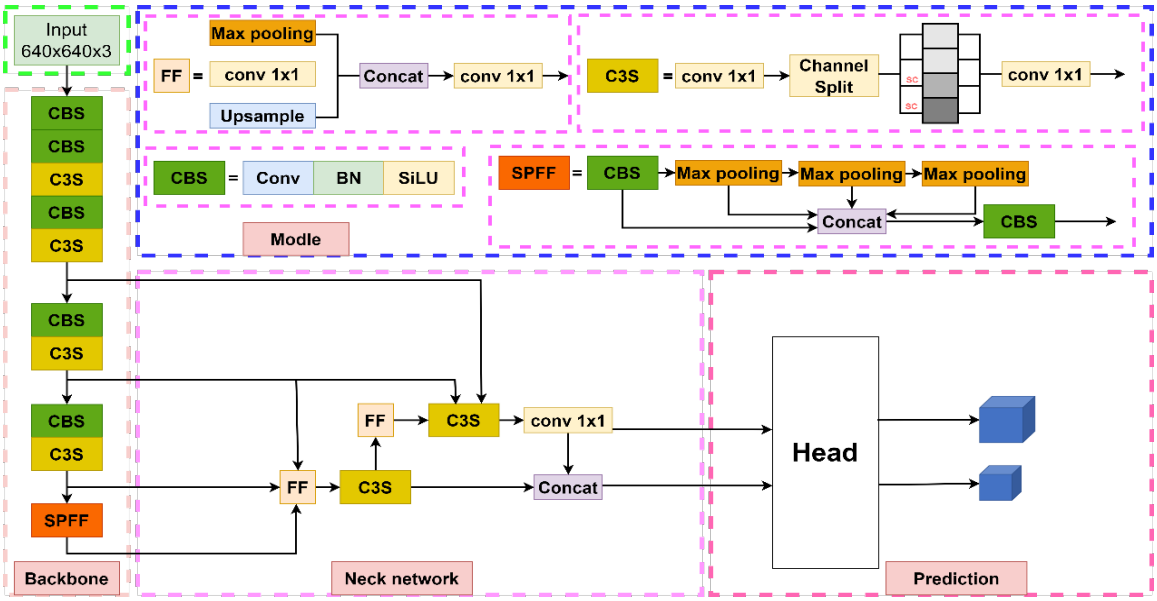


Figure 2. Structure diagram of improved YOLOv5s. Note: CBS is Conv+BN+Silu, Conv is convolution operation, BN is normalization operation, Silu is sigmoid weighted linear combination activation function, SPFF is fast spatial pyramid pooling structure, and Concat is feature fusion function.

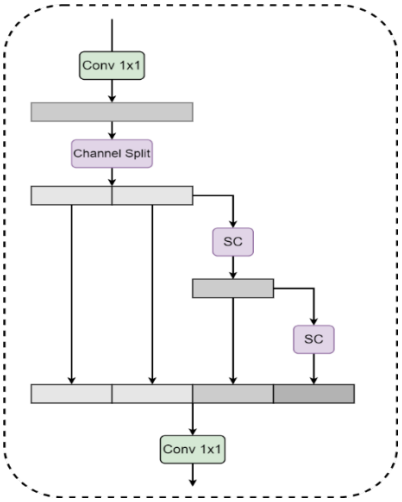


Figure 3. C3S Module Structure.

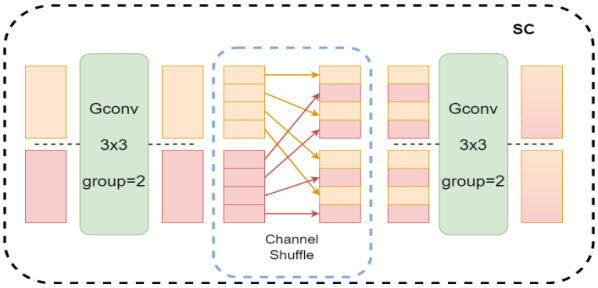


Figure 4. SC Module Structure.

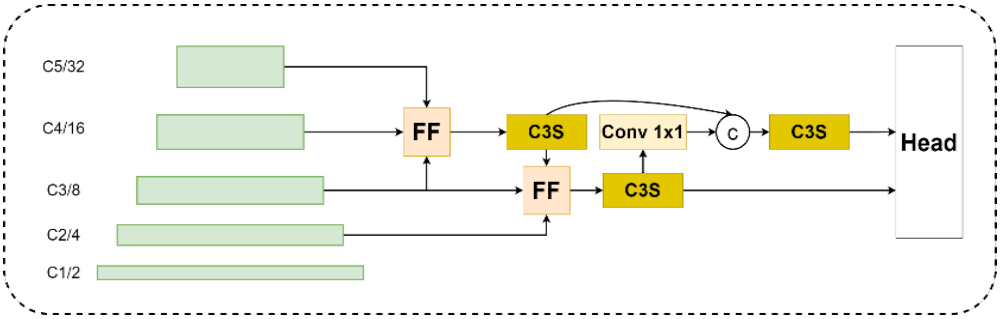


Figure 5. 3-Feature fusion dual head detection.

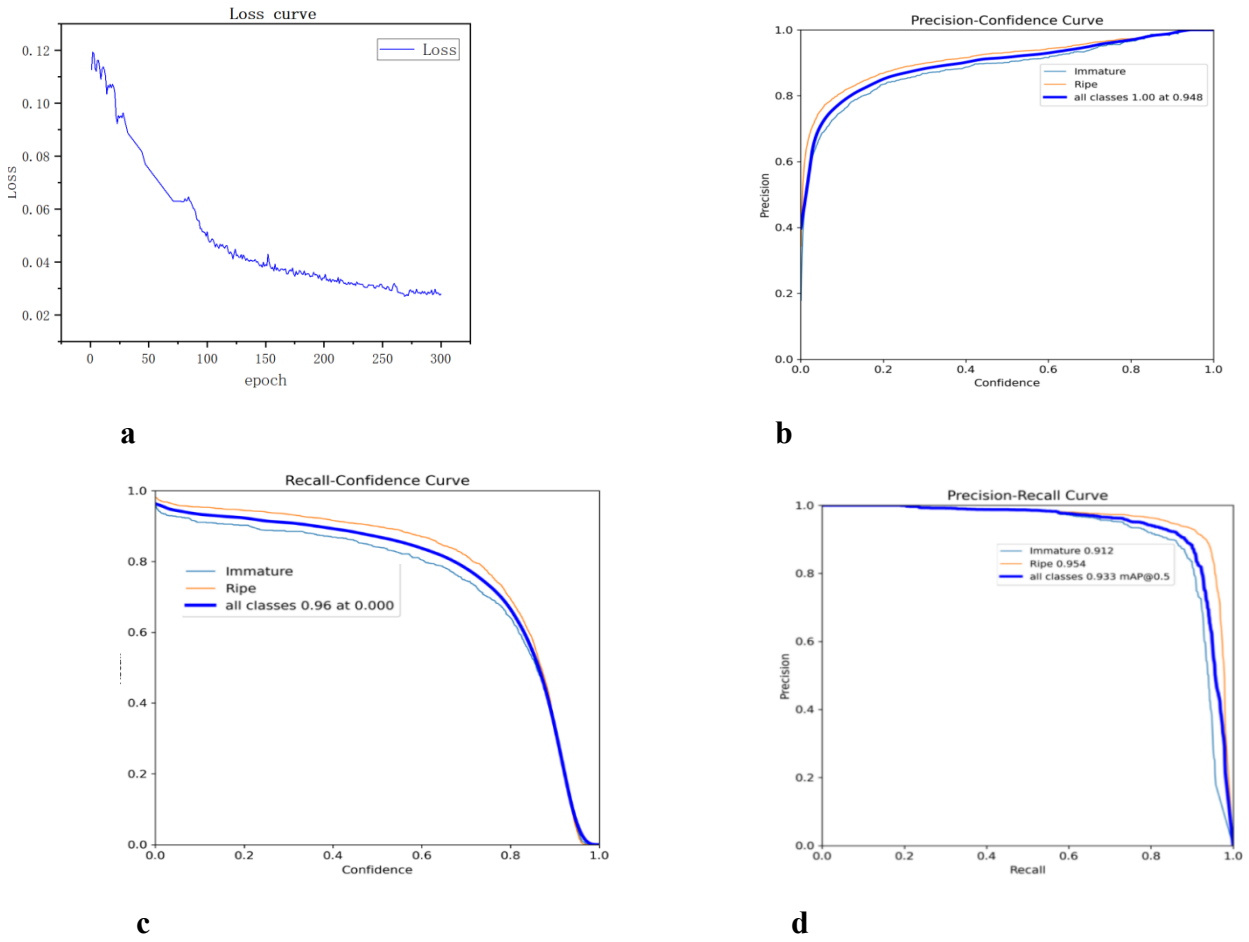


Figure 6. Loss function P, R, P-R change curve during training. a) Loss curve; b) Accuracy curve; c) Recall rate curve; d) Accuracy Recall Curve.

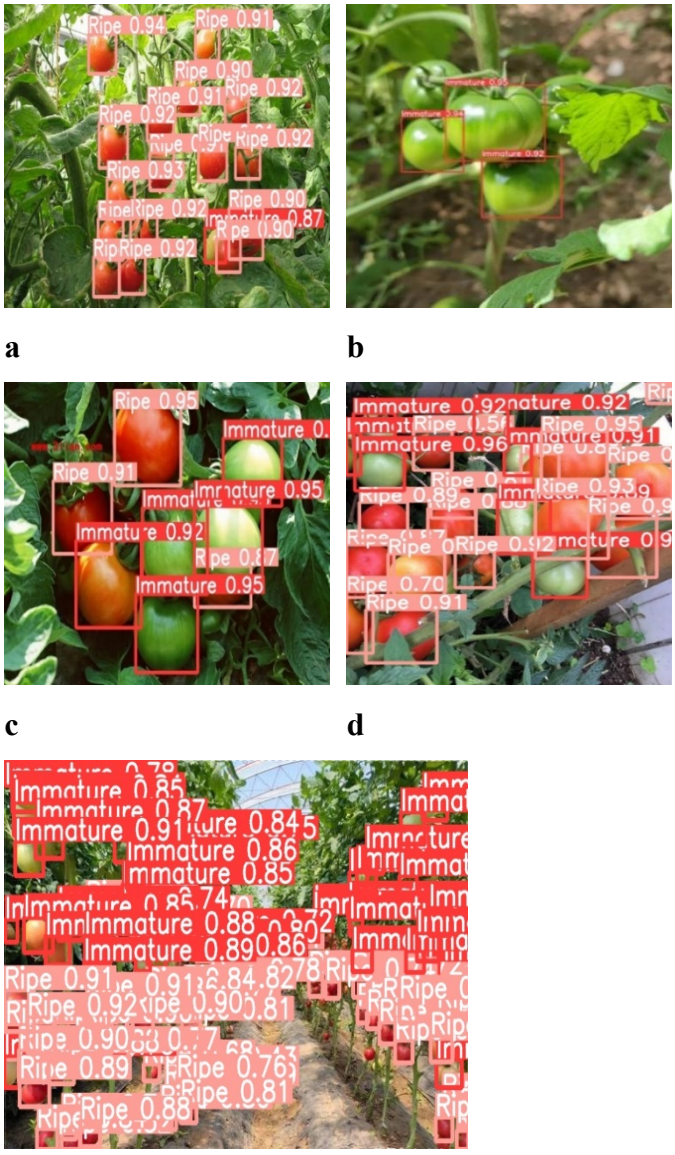
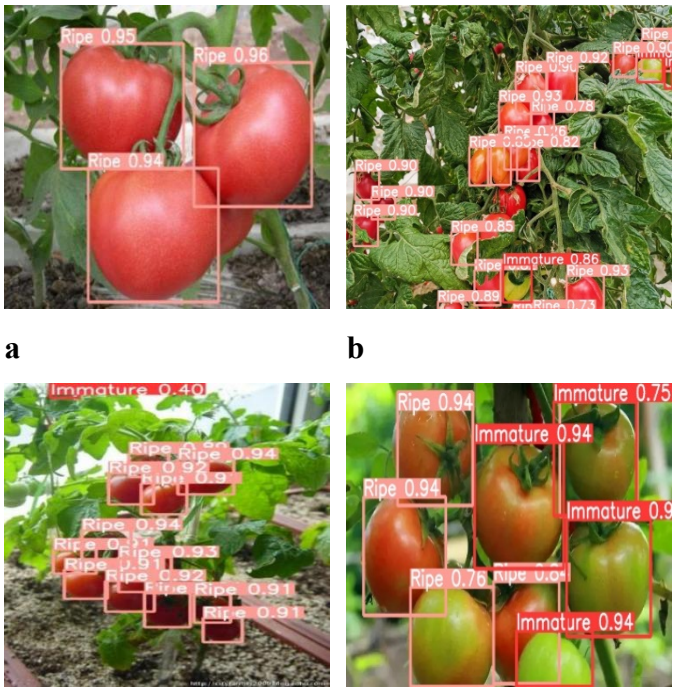


Figure 7. Effect of tomato fruit detection. a) Effect of mature tomato detection; b) Detection effect of immature tomatoes; c) Occlusion between fruits; d) Fruit dense growth; e) Coexistence of Two Maturity Tomatoes at Wide Angle.



a b

c d

Figure 8. Example diagram of missed detection and false detection. a) Misinspection caused by obstruction between fruits; b) Missing detection caused by dense fruits growth; c) Misdetection caused by similarity in leaf characteristics with immature tomatoes; d) Misdetection caused by similarity in characteristics between ripe tomatoes and tomato during the color transition period.

Table 1 Ablation experiment performance comparison.

Model	Accuracy (%)	Recall (%)	mAP0.5 (%)	Speed (fps)	Floating point operation (GFLOPs)
YOLOv5s	94	94.1	92.8	82	15.8
YOLOv5s1①	92.5	90.5	93.2	98	9.2
YOLOv5s2②	93.7	92.8	93.6	93	11.8
YOLOv5s3	94.8	96	93.3	106	8.3

Note: ○,1Add C3S module; ○,2Add an FF module (3 feature fusion dual head detection) with an average accuracy of each category when the $mAP_{0.5}$ IoU threshold is 0.5; fps represents the number of frames processed per second, which is used to measure the detection speed of the model; GFLOPs means Giga Floating point Operations Per Second, which is the number of floating-point operations that occur 1 billion times per second.

Table 2 Detection performance of different models.

Model	P (%)	R (%)	mAP0.5 (%)	Model parameter quantity (M)	Detection speed (ms)
YOLOv5s3	94.80	96.00	93.30	3.02	9.4
YOLOv5s	94.00	94.10	92.80	7.01	12.2
YOLOv5m	96.20	97.00	93.10	21.17	14.1
YOLOv5l	90.30	96.00	81.10	45.56	18.4
YOLOv5x	96.30	96.00	94.30	86.71	29.7
YOLOv5n	94.30	97.00	89.20	4.08	11.1
YOLOv4	91.50	88.83	89.32	94.83	12.7