Research on Crop Detection Algorithm Based on Improved YOLOv7

Xiaoqi Shi
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: shixiaoqi713@163.com

Xin Ye
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, Shaanxi, China
E-mail: yexin@xatu.edu.cn

***Abstract*—In the field of crop target detection, traditional target detection algorithms are often difficult to achieve satisfactory accuracy due to factors such as dense distribution of species and poor imaging quality, which brings many inconveniences and challenges in practical agricultural production applications. To address this situation, the study introduces an enhanced YOLOv7 algorithm, incorporating the attention mechanism, with the objective of substantially elevating the overall performance in crop target detection tasks. The improved algorithm can more accurately focus on the key features of crops by cleverly incorporating the attention mechanism, effectively filtering out the interference of complex background and noise, so as to achieve more accurate recognition of various crops. After a large amount of experimental data verification, the improved algorithm can achieve an average recognition accuracy of 80% for a variety of crops, with an average accuracy of 75%, and the highest recognition efficiency is as high as 91% in the detection of some specific crops. In contrast to other prominent crop target detection algorithms, the refined algorithm presented in this paper exhibits remarkable performance benefits. Notably, its target detection efficacy is highly significant, enabling swift and precise identification of crop species.**

Keywords-Target Detection; Attention Mechanism; YOLOv7; Crop Species Recognition

I. INTRODUCTION

Our country is a country dominated by agriculture in historical records, and fruits and vegetables have an indispensable position in the everyday lives of our populace. In recent years, China's fruit and vegetable industry has developed rapidly, with the development of domestic vegetable farming and the continuous introduction of foreign fruits and vegetables. Crops in quantity, quality and category also meet the growing demand of urban and rural residents, which makes China's fruit and vegetable production and sales are also increasing year by year. However, in recent years, more and more people have been farther and farther away from agricultural production, resulting in a lack of knowledge about common crops. The conventional approach to identifying crop species primarily involves manual visual inspection, which is not only inefficient but also constrained in accuracy by the inspectors' experience and skill levels.

In recent years, China's fruit and vegetable industry has undergone swift development, accompanied by a steady influx of exotic fruits and vegetables， in terms of quantity, quality and variety of categories to meet the growing needs of urban and rural residents, a large number of fruit and vegetable products break through the original cognition of the people, and therefore urgently need related technology to help people quickly identify the relevant products [1]. Due to the swift advancements in machine vision and artificial intelligence, AI technology has progressively integrated into various facets of production and daily life. Intelligent classification and identification of fruits and vegetables by means of deep learning machine vision is an ideal way of processing [2].

In this study, the model's foundational architecture is chosen to be the YOLOv7 algorithm, and to augment its recognition precision, the SE-Net attention mechanism is integrated into its main network structure. Following these enhancements, the YOLOv7 model's performance in crop recognition tasks has

undergone notable improvement in accuracy, and all the experimental results meet the expected goals. These improvements effectively validate the initial idea of performance enhancement of the model.

II. Related Work

Foreign research in fruit and vegetable identification began at an earlier period, foreign scholars used LiDAR technology and deep learning to predict vegetable crop growth, combined with LiDAR (laser radar) technology, formulated a deep learning-based prediction model, capable of estimating the height and canopy size of vegetable crops. On an experimental farm at Bangalore Agricultural University in India, data pertaining to the growth cycles of tomatoes, eggplants, and kale were gathered across five distinct time points spanning a specific period. The research team at Bangalore Agricultural University, India, used a terrestrial laser scanner to acquire LiDAR point clouds and integrated a hybrid deep learning architecture that merges Long Short-Term Memory Networks (LSTMs) with Gated Recurrent Units (GRUs).to make predictions [3]. The deep learning model exhibited an approximately 80% accuracy in anticipating structural parameters at the plant level for various stages of crop growth in advance. Specifically, the hybrid model demonstrates efficacy in forecasting canopy area, with height prediction errors ranging from 5% to 12%, and a balanced occurrence of both over- and underestimation biases.

Domestic research in the direction of fruit and vegetable identification started relatively late, and most of them only studied and improved the algorithm. In 2016, Zeng Weiliang et al. designed a fruit and vegetable recognition system for smart refrigerators through convolutional neural networks, which is based on the improved LeNet-5 algorithm, on which the ReLU activation function is used and the Droupt technique is used. It was tested on a dataset with 15 categories of fruits and vegetables and a total of 2633 images of fruits and vegetables, and the experimental results obtained an accuracy of 83.4% [4]. In addition, in 2020, by Cheng Shuai, Li Yanling, Si Haiping, and Sun Changxia, the network parameters were optimally adjusted based on the DenseNet121 network. It was trained and tested on a kind of dataset with only five types of crops and compared with the classical VGG16 model and ResNet50 model on the same dataset, and the evaluation results indicated an enhancement in the network's recognition rate by 1.1% and 6.9%, when compared to the two respective algorithms [5].

Other scholars have used deep learning algorithms for feature extraction, spot segmentation, and detection and recognition of different disease classes on crop leaves. To identify crop leaf diseases, they employed diverse deep learning techniques, including convolutional neural networks (CNNs) and support vector machines (SVMs), among others. Through the construction of a deep learning model, followed by its training and optimization processes, high accuracy recognition of crop diseases was achieved [6]. The deep learning-based crop disease recognition method is faster and simpler than the traditional recognition method, and the recognition accuracy is improved. Some studies have shown that deep learning-based plant disease classification and recognition methods can achieve 91% to 98% recognition accuracy.

III. Technical model

A. YOLOv7 Algorithm

YOLOv7 is a network model in the YOLO family introduced in recent years. This model is currently the YOLO model with the fastest inference speed and best recognition results on the PASCAL VOC dataset. When compared to other target detection models, it surpasses in both speed and accuracy, thereby fulfilling the requirement for prompt and precise identification of crop species in natural settings. However, although the original YOLOv7 algorithm has already attained high detection accuracy, there remains potential for enhancement to mitigate the influence of similar features, target scale and shape variations on the detection accuracy. YOLOv7 proposes a new network architecture called ELAN (Efficient Layer-wise Adaptive Network), which focuses on high efficiency. The ELAN framework facilitates efficient learning and convergence in deeper networks by regulating the gradient paths, whether they are the shortest or longest [7].

YOLOv7's loss function comprises components for coordinate error, confidence in target prediction, and classification accuracy. The matching strategy employs the SIMOTA method, which obtains anchor frames by k-means clustering and positive sample expansion. The network architecture of YOLOv7 consists of the CBS (Conv, BatchNorm, Silu) module, the CBM module, the REP module, the MP module, the ELAN module, the ELAN-W module, the UP Sample module and SPPCSPC modules. These modules work together to achieve efficient feature extraction and target detection [8].

YOLOv7 uses Leaky ReLU (Rectified Linear Unit with Leakage) as an activation function [9], which solves the problem of the traditional ReLU function having zero gradient in the negative region.

It is due to the above characteristics that the YOLOv7 algorithm has been widely used in application scenarios where the demand is for high speed and high accuracy, and it also meets the criteria for speed and precision in detection in this crop species identification system. Therefore, this study utilizes YOLOv7 as the foundational detection model and further refines it. The YOLOv7 network architecture is shown in Figure 1.

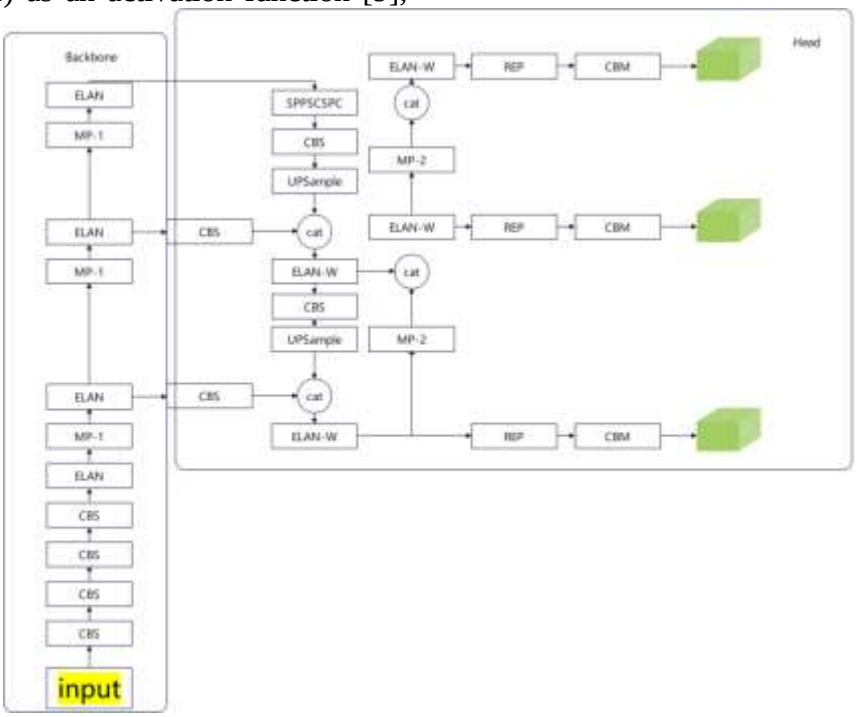

Figure 1. YOLOv7 network architecture diagram

In order to further improve the detection accuracy of the model, this study introduces an enhanced algorithm built upon the YOLOv7 framework, specifically, the method of adding an attention mechanism to the Backbone network to enhance the precision of detection using the YOLOv7 algorithm [10], by comparing the model detection accuracy before and after the improvement.

The Backbone network in the YOLOv7 algorithm is mainly composed of two major modules, Multi_Concat_Block module and Transition Block module [11].

1) Multi_Concat_Block Module

There are four feature layers that perform the final feature stacking in this module, the 0th bit labeling, the 1st bit labeling, the 3rd bit labeling, and the 5th bit labeling. Bit 0 does not operate, bit 1 performs 1 convolution, bit 3 performs 3 convolutions, bit 5 performs 5 convolutions, and after feature stacking the features are integrated by one convolution, in this module, the main

operation performed is bitwise convolution. As shown in Figure 2.

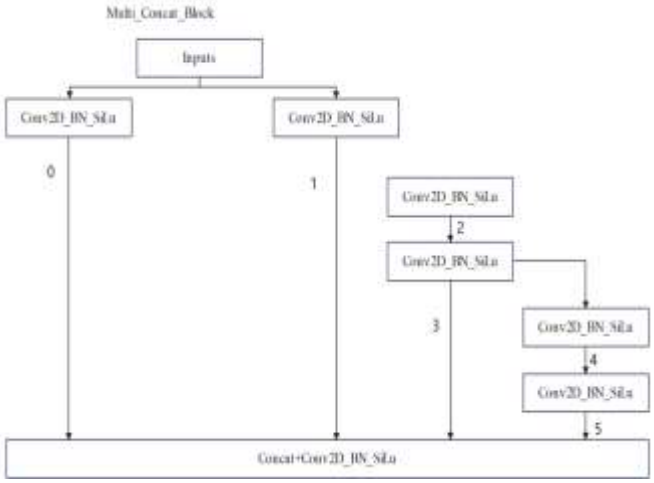

Figure 2. Schematic diagram of Multi_Concat_Block module

2) Transition Block Module

The module consists of two main branches, the left branch of the input data into a step for the maximum pooling, and then through a convolution to adjust the number of channels from 1024 to 512. the right branch of the input data into a convolution and then through a convolution kernel size of the step for the convolution of the extraction of features. Refer to Figure 3 for illustration.

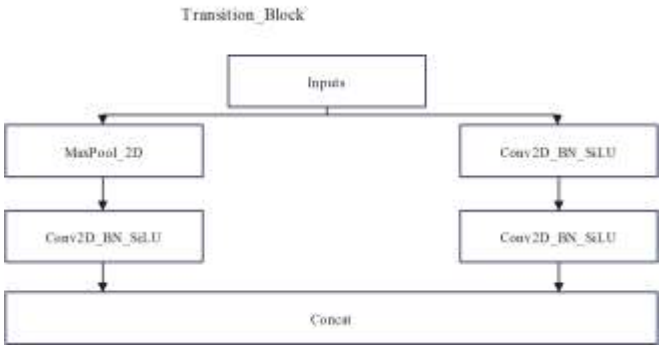

Figure 3. Schematic diagram of Transition Block module

3) FPN Enhanced Feature Extraction

The role of SPPCSPC module is to increase the sensory field, the module is mainly divided into two branches, the left branch of the first four branches were 5, 9, 13, 1 max-pooling [12], this process enables four sensory fields to distinguish between large and small objects. For example, there are different sizes of fruits and vegetables in a photo, their sizes are not the same, after this module will be able to better distinguish between small and large targets, SPPCSPC structure sketch depicted in Figure 4.

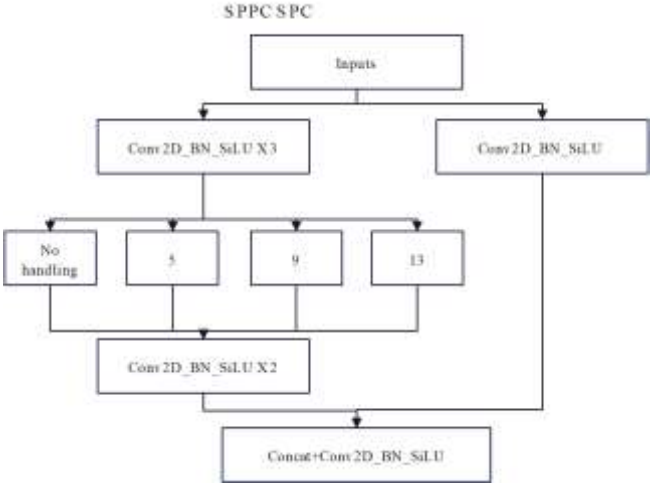

Figure 4. Sketch of SPPCSPC structure

The following are the steps for feature fusion at the FPN layer.

Step1. Using SPPCSPC for feature extraction to operate on the bottom feature layer can improve the sensory field of YOLOv7 algorithm, making YOLOv7 algorithm can be more accurate in recognizing objects with different sizes within the picture.

Step2. Conduct 1×1 convolution to adjust the channel, and then perform up-sampling operation to combine with the feature layer after one convolution of the middle and lower feature layers, and then use Multi_Concat_Block module for feature extraction.

Step3. Perform 1×1 convolution to adjust the channel for the feature layer obtained in the second step, then perform up-sampling and combine it with the feature layer of the middle layer after one convolution, and then use the Multi_Concat_Block module for feature extraction.

Step4. The feature layer obtained in the third step is down sampled by the Transition Block module once, and then down sampled and stacked with the feature layer obtained in the second step,

after which feature extraction is performed using the Multi_Concat_Block module.

Step5. The feature layer obtained in the fourth step is down sampled by the Transition Block module once, and then stacked with the feature layer obtained in the first step after down sampling, and then the Multi_Concat_Block module is used for feature extraction.

Up to this point, three enhanced feature layers are obtained, which are the feature layers obtained in the third, fourth and fifth steps, respectively. The feature pyramid can complete the feature fusion operation between feature layers of different shapes, which is helpful to extract better features [13]. The feature layer shape change diagram is shown in Figure 5.

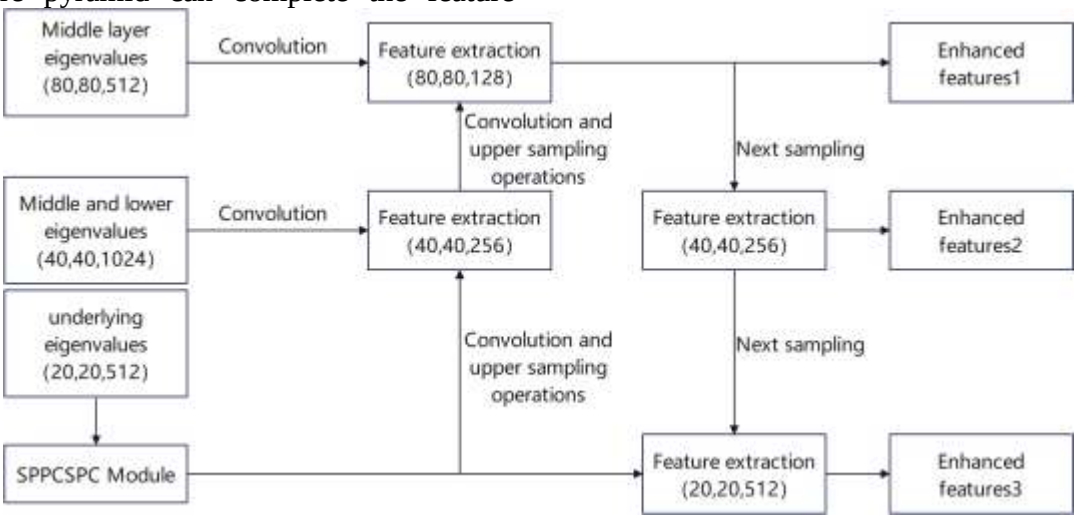

Figure 5. Feature layer shape change map

Add SE-Net attention mechanism, its full name is Squeeze-and-Excitation Networks, YOLOv7 algorithm is still a kind of algorithm based on the improvement of CNN in essence [14], and SE-Net attention mechanism is a kind of model that introduces the attention mechanism in the convolutional neural network.

The attention mechanism in SE-Net is mainly realized through SE Block, the specific procedures are outlined below.

Step1. Squeeze Compression. The SE-Net attention mechanism employs global average pooling to reduce each channel's feature maps to a single value, capturing global spatial information across channels. The dimensionality of this step is varied as $(C,H,W)\to(C,1,1)(C,H,W)\to(C,1,1)$.

Step2. Excitation. The SE-Net attention mechanism uses two fully connected (FC) layers to learn the relationships between channels, generating a weight for each channel. The first fully-connected (FC) layer performs a dimensionality reduction and the second fully-connected (FC) layer reverts to the original dimensions and an activation function is applied to guarantee positive weights summing to one. The dimensionality change for this step is $(C,1,1)\to(C,1,1)(C,1,1)\to(C,1,1)$.

Step3. Scale deflation. Multiply the weights obtained from the motivation step by the original feature map, the dimensionality change of this step is $(C,H,W)\times(C,1,1)\to(C,H,W)(C,H,W)\times(C,1,1)\to(C,H,W)$

The attention mechanism is introduced in the three intermediate layers, the lower middle layer and the bottom feature layer positions in the Backbone network mentioned earlier, and the specific introduction position of the attention mechanism is shown in Figure 6.

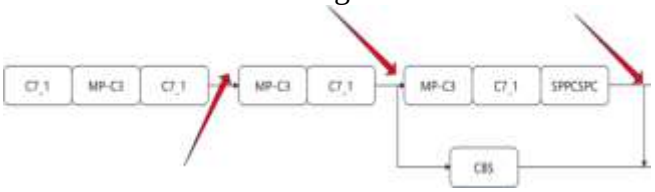

Figure 6. Map of the location of the introduction of the attention mechanism

IV. Experiment and Analysis

B. Experimental Content

Target detection of crop species, in the establishment of the data set, take the fruit and vegetable photos obtained on the Internet to build their own data set, the pictures include multiple angles, as well as different lighting conditions and increase the interference term and other forms of sampling, and then organize the experimental data set suitable for the needs of the experiment, and ultimately a total of seven common vegetables as well as seven types of fruits to form a data set, a total of 14 different types of Crops. To assess the method's performance across various training iterations in this study, experimental demonstrations were conducted on this data set.

Compared to traditional image recognition methods, the YOLO algorithm introduced in this study demonstrates superior results in detecting crop species targets The refined YOLOv7 method achieves notable enhancements in processing speed as well as detection precision.

The environment for this experiment is the server operating system is Ubuntu 20.04, the number of CPU cores is 8 cores, RAM is 15G, and the GPU is GeForce RTX 2080 Ti with 11G of video memory. To speed up the training time, the GPU is used for acceleration, and the code is written using the Python 3.9 programming language. The constructed dataset was labeled using Labellmg software. SSH connection was used to connect the local machine to the cloud server, and Xftp7 software was used to upload project code and files to the cloud server, as well as to download the trained models.

C. Experimental Process

The roadmap for the realization of the experiment is shown in Figure 7.

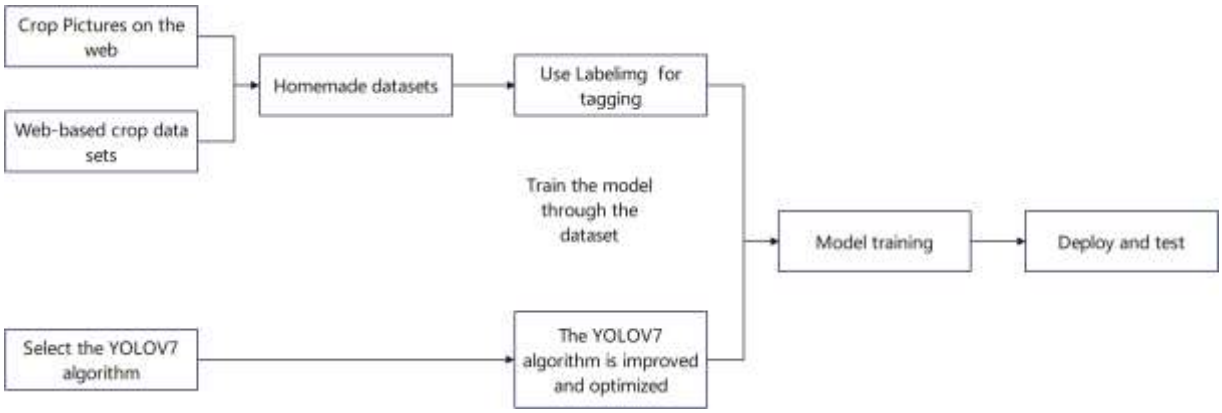

Figure 7. Roadmap for system realization

This paper uses a homemade dataset with specific types and quantities of fruits and vegetables as shown in Table 1.

TABLE I. Table of fruit types and corresponding number of pictures

Name and number of vegetables	Name and number of fruit
Cabbage（200）	Apple（200）
Capsicum（200）	Banana（200）
Carrot（200）	Pear（200）
Cauliflower（200）	Pineapple（200）
Corn（200）	Pomegranate（200）
Eggplant（200）	Grapes（200）
Cabbage（200）	Apple（200）

In the preprocessing stage, the homemade dataset was image labeled using Labellmg image labeling software, and the dataset was allocated into training, validation, and testing subsets in an 8:1:1 ratio.

D. Experimental Results

The models before and after the improvement are trained in the same dataset and experimental environment framework in the experiments, and whether the improvement has improved the accuracy is obtained by comparing the detection accuracy of the two models.

The confusion matrix obtained here is a 14×14 matrix as there are 14 categories in total.

In the matrix, rows signify actual categories while columns represent predicted categories. Diagonal elements indicate correct classifications, whereas off-diagonal elements signify misclassifications in the matrix.

The confusion matrix plot obtained at the end of training is shown in Figure 8. Where banana, cabbage, pepper, cauliflower, corn, pear, pineapple and pomegranate have a value of 1.00 on the diagonal, which indicates that the model is better trained for these eight types of crops, while the other types of crops have the problem of prediction error during model training, where the grapes have the largest prediction error of 0.50. this indicates that this model has a lower prediction for grapes.

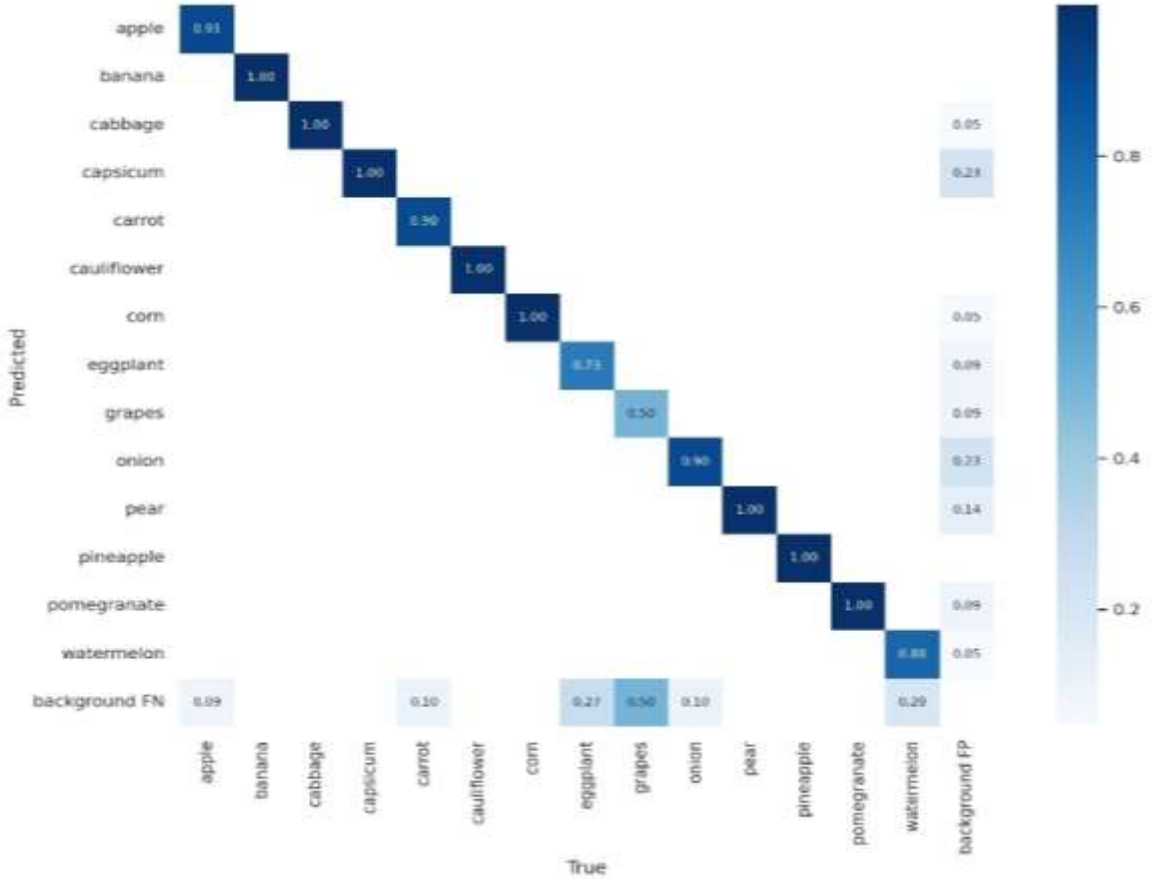

Figure 8. Confusion matrix diagram

The reconciled mean of checking accuracy and recall is defined as F1, with a maximum of 1 and a minimum of 0. Typically, a greater F1 score indicates superior model performance, as illustrated in Formula (1).

$$F1 = \frac{P \times R}{P + R} \times 2 \quad (1)$$

Figure 9 displays the F1 images, with the horizontal axis depicting various thresholds and the vertical axis showing the corresponding F1 scores, and the altered image indicates that the F1 scores of all the species were maximized in the interval from 0.80 to 0.417.

The experiment's P_curve graph illustrates the correlation between accuracy and confidence, a prevalent method for visualizing target detection outcomes. The P-curve plot is advantageous as it facilitates the assessment of detector performance and the determination of an appropriate threshold by depicting accuracy curves across varying confidence levels. Figure 10 illustrates this concept. The plot's horizontal axis depicts the confidence level, while the vertical axis represents the accuracy rate. Each thin line on the plot

corresponds to the accuracy curve of a specific category, and the thick line represents the mean precision curve across all categories. The confidence level, if too high, may miss some real samples with low determination probability. From Figure 10 below, we can see that at a confidence value of 0.942, the model in this paper achieves perfect accuracy (i.e., no false alarms) for all categories, which is in line with the trend that this figure should have, and the results are better.

The PR curve of the experiment represents the relationship between the precision and the recall. The PR curve serves to demonstrate the model's performance across varying thresholds by plotting the precision and recall rates as the decision thresholds change, and the performance of the model is better indicated when the curve is positioned closer to the upper right corner. The PR curve of the training results of the experiments in this paper is shown in Figure 11, on the PR curve, the recall rate is plotted along the horizontal axis, while the precision rate is represented on the vertical axis. AP denotes the area under the curve, mAP refers to the average of the average precision (AP) of all target categories in the model, and the threshold used to classify IoU as a positive or negative sample is denoted by the number following mAP@. mAP@0.5 denotes the mean mAP for which the threshold is greater than 0.5. The calculation is shown in Formula (2).

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \qquad (2)$$

Where N in the formula indicates the number of all categories, in this experiment N=14, which is indicated as the mean precision of the i-th category, the mean precision value can be obtained by adding the accuracy values of its 14 categories and dividing by 14. It can be seen that the mAP@0.5 of target detection of this paper's model is 0.822, and the obtained experimental results align with the anticipated outcomes.

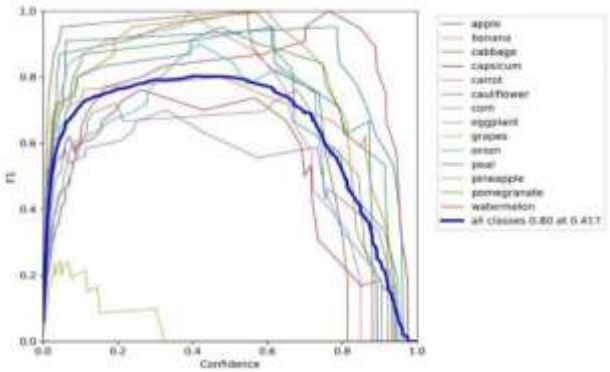

Figure 9. F1 score graph

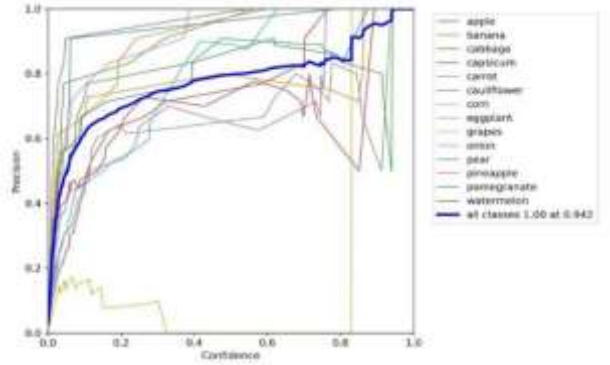

Figure 10. P_curve

The top right corner of Figure 11 below shows the 14 crop categories and their respective accuracy values during training.

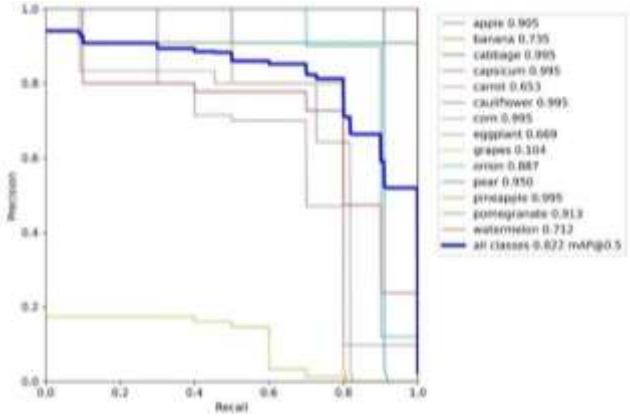

Figure 11. PR Curve

The R_curve plot of the experiment represents the relationship between recall and confidence, and the function of the R_curve plot is to understand the model's ability to recognize positive samples under different conditions by plotting the recall curves under different confidence levels. As shown in Figure 12. The

horizontal coordinate of the graph is the confidence level and the vertical coordinate is the recall, where the thin line represents the recall curve for each category and the thick line represents the average recall curve for all categories. Ideally, when the confidence threshold is 0, all the detected frames are retained, and thus the recall should be 1. However, in practice, factors such as errors in the target detection algorithms and noise may cause some detected frames to be incorrectly filtered out, which results in a recall that is not 1.

At lower confidence levels, categories are detected more inclusively. Specifically, when the confidence level is set to 0, the average recall across all categories achieves 0.95, surpassing experimental benchmarks and fulfilling requirements.

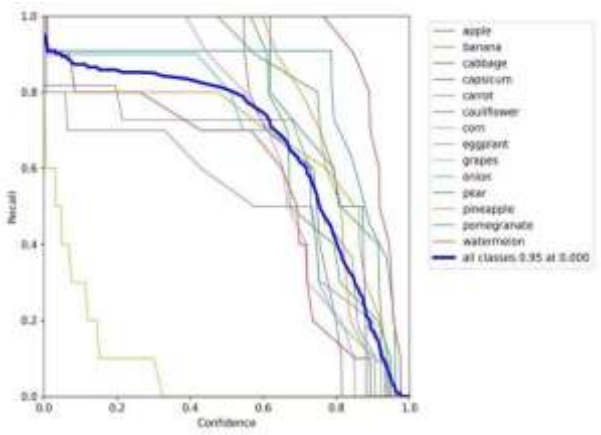

Figure 12. R_curve

Figures 13 and 14 present the experimental outcomes, depicting results for various crops.

Figure 13. Apple experiment result

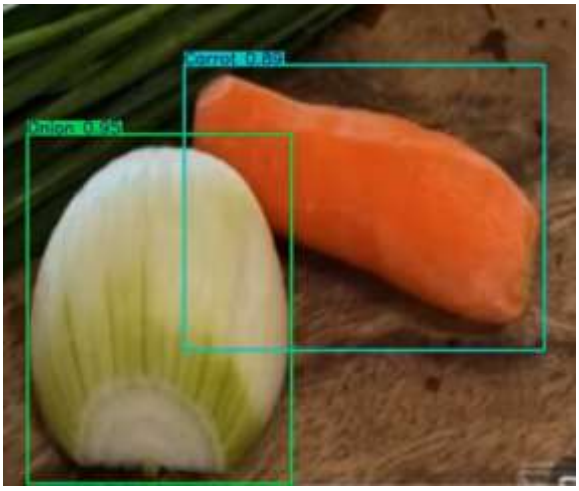

Figure 14. Results of Onion and Carrot experiments

After detecting a large number of images, the detection accuracy values obtained from each detection are recorded and the resulting detection accuracy is shown in Table 2.

TABLE II. COMPARISON TABLE OF DETECTION ACCURACY

Type	Evaluation metrics		
	Detection Times	mAP/% (Pre-improved)	mAP/% (Improved)
Apple	30	0.79	0.85
Banana	30	0.74	0.77
Pear	30	0.79	0.81
Pineapple	30	0.81	0.79
Pomegranate	30	0.68	0.68
Grapes	30	0.50	0.57
Watermelon	30	0.73	0.78
Cabbage	30	0.89	0.91
Capsicum	30	0.55	0.58
Carrot	30	0.83	0.89
Cauliflower	30	0.55	0.64
Corn	30	0.46	0.45
Eggplant	30	0.74	0.71
Onion	30	0.88	0.95

By summarizing the data in Table 2 above, it can be seen that the improved model is higher than the pre-improved model in all the other 11 categories, although three categories are lower

than the pre-improved one in terms of detection accuracy.

V. CONCLUSIONS

For each crop, the average value of the detection accuracy for each category is different, both before and after improvement, which is due to the fact that when training the model, the detection of some categories is not very satisfactory due to some external factors, such as incomplete image annotation when building the dataset on its own, and thus the detection of some categories is not very satisfactory. In this paper, a fruit and vegetable dataset suitable for this study is constructed by collecting and organizing the data, including seven kinds of vegetables commonly found in supermarkets and farmers' markets as well as seven kinds of fruits, with a total of 14 categories and 3220 images. In this experiment, the YOLOv7 algorithm serves as the foundational network for the entire model, and the SE-Net attention mechanism is added to its backbone network to improve the accuracy of the model. It makes up for the past defects such as low accuracy of crop recognition, etc. Using the improved model to train the dataset, several sets of experiments were conducted, and the results of the experiments were analyzed and discussed to conclude that the improved approach of this study has a positive effect on this experimental dataset.

The improved YOLOv7 model achieved an average accuracy of 80% and an average precision of 75% in crop species recognition. All the experimental results met the envisioned expected values. In future research, further lightweighting studies of the network model will be conducted to enhance the model's detection rate.

REFERENCES

[1] ZouJunRong. 2022 Fruit and Vegetable Industry Development Outlook and Market Trend Research and Analysis. https://www.chinairn.com/scfx/20220715/121648626.shtml, 2022-7-15).

[2] Pan Mei. Application of image recognition in fruit and vegetable classification and recognition[J]. Modern Agricultural Science and Technology,2021(16):257-259.

[3] J, R., Nidamanuri, R.R. Deep learning-based prediction of plant height and crown area of vegetable crops using LiDAR point cloud[J]. Scientific Reports,14,14903(2024).

[4] Zeng Wei-liang, Lin Zhi-xian, CHEN Yong-shan. Research on fruit and vegetable image recognition for smart refrigerator based on convolutional neural network[J]. Microcomputer and Applications,2017,36(08):56-59.

[5] Cheng Shuai, Li Yanling, SI Haiping, et al. Research on automatic crop species identification algorithm based on convolutional neural network[J]. Henan Science,2020,38(12):1908-1914.

[6] B. V S, Harshad B, Vijay D. Hunger games search based deep convolutional neural network for crop pest identification and classification with transfer learning[J]. Evolving Systems,2022,14(4):649-671.

[7] Zhao Pengfei, Qian Mengbo, Zhou Kaiqi, et al. Improvement of sweet pepper fruit detection in YOLOv7-Tiny farmland environment[J]. Computer Engineering and Application,2023,59(15):329-340.

[8] Wang Yihan. Research on citrus fruit recognition and localization method in natural environment based on improved YOLOv7[D]. Sichuan Agricultural University,2023.

[9] Wu Jie, Shi Lei, Zhang Zhi-An. Research on pest image recognition and classification method based on deep learning[J]. Computing Technology and Automation,2023,42(01):166-173.

[10] Luo Tonglan. Research on potato defect detection based on improved YOLOv7[D]. Ningxia University,2023

[11] Pang Haitong. Research on intelligent identification technology of orchard pests based on deep learning[D]. Zhejiang University,2021.

[12] Fang Si-Wen. Research on apple localization and identification technology in complex environment based on YOLOv7[D]. Henan Agricultural University,2024.

[13] Jian Huang, Gang Zhang. A review of target detection algorithms for deep convolutional neural networks[J]. Computer Engineering and Application,2020,56(17):12-23.

[14] Xu Qiu. Research on transmission line bird damage detection and hazardous bird species identification based on YOLOv7[D]. Nanchang University,2024.