IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 November 2023, accepted 20 November 2023, date of publication 23 November 2023,
date of current version 29 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3336562

RESEARCH ARTICLE

Revolutionizing Agriculture: Real-Time Ripe Tomato Detection With the Enhanced Tomato-YOLOv7 System

JUN GUO[1], YUE YANG[1], XINYAN LIN[1], MUHAMMAD SOHAIL MEMON[2,3], WEI LIU[1], MEIQI ZHANG[1], AND ENHUI SUN[1]

[1]School of Automotive Engineering, Yancheng Institute of Technology, Yancheng, Jiangsu 224051, China
[2]Department of Farm Power and Machinery, Faculty of Agricultural Engineering, Sindh Agriculture University, Tando Jam 70060, Pakistan
[3]School of Agricultural Engineering, Jiangsu University, Zhenjiang 212013, China

Corresponding author: Yue Yang (ycit_yy@163.com)

This work was supported in part by the Yancheng Institute of Technology, Yancheng, Jiangsu, China, through the School-Level Research Projects under Grant 154203635.

ABSTRACT Traditional agricultural practices of hand-picking ripe tomatoes are labor-intensive and inefficient for large-scale harvesting. To address this, we propose an innovative approach using the YOLOv7 algorithm for ripe tomato detection, enabling robotic arms to perform the picking. However, the occlusion of tomatoes in the field often leads to unclear target features, causing false or missed detections. So it is worth studying and this paper proposes a tomato detection method based on improved YOLOv7. The novelty is shown below. First, a new structure called ReplkDext is redesigned to increase the receptive field. ReplkDext is introduced before the last layer of CBS in the backbone. Secondly, to overcome the problem of low FLOPS caused by frequent access to memory in traditional neural networks, the head structure of YOLOv7 is redesigned. By using FasterNet to optimize the structure between Concat and CBS in the head, FasterNet makes the model balance between running speed and detection accuracy. Finally, to improve the ability of convolution, ODConv is added after the last ELANN-2 structure in the Head layer. ODConv improves the feature extraction ability of small targets and obtains more feature information about ripe tomatoes. Experiments show that compared with YOLOv7, Map@.5 of Tomato-YOLOv7 has increased by 1.3%. The model is overall better than other models. The overall contribution of the Tomato-YOLO model is to provide important insights into agricultural product detection and provide a theoretical basis for automated tomato harvesting in orchards.

INDEX TERMS Tomato, improved YOLOv7, target detection, occlusion, missed detection, map.

I. INTRODUCTION

In orchards, picking ripe tomatoes is usually done by hand. This way wastes a lot of time and human resources [1], [2]. To solve the problem about low efficiency of manual picking, this paper applies machine vision to picking tomatoes. Machine vision is mature and used in traditional and modern industries, such as express sorting [3] and road vehicle detection [4]. Machine vision is rarely used in agriculture, so it is of great significance to apply machine vision to agriculture.

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenhua Guo.

Object detection is an important task in machine vision. Traditional object detection has the problems of long detection time and poor robustness [5]. On the contrary, object detection in deep learning has the advantages of short detection time and high detection accuracy, which greatly meets the requirements of real-time detection in complex environments. Currently, the mainstream deep learning model is YOLOv5, YOLOv6 [6], YOLOv7 [7], and YOLOv8. In ripe tomato detection, due to mutual occlusion between tomatoes, the detector is easily disturbed by occlusion, and a wrong prediction frame is given. Since YOLOv7 has the advantages of being balanced in speed and accuracy, this paper proposes

an improved YOLOv7 to improve the ability of YOLOv7 to detect ripe tomatoes. The improved algorithm has theoretical guiding significance for detecting small targets or occlusions, and has important practical significance for applying agricultural target detection.

In tomato detection, the detection object has few effective pixels, a small scale, and lack of feature expression ability due to mutual occlusion or small targets [8], [9], [10]. Therefore, some tomatoes may not be detected in actual tomato detection. To solve this problem, this paper optimizes the structure of the YOLO model to improve the feature extraction capability and accuracy of the model. In experimental results, certain indicators are commonly used to assess whether the model's performance has improved, such as precision (P), recall (R), mean average precision (MAP), frames per second (FPS), parameters (Params), and floating-point operations (FLOPs) [11].

To sum up, the novelty of the current work in this paper is to optimize the ELANB and ELANN structures in the YOLOv7 model. First, this paper redesigns a new structure called ReplkDext to increase the receptive field of the model's backbone network. Second, the FasterNet structure is added to enable the model to strike a balance between speed and detection accuracy. Finally, adding ODConv to the last small target detection part in the head layer can effectively improve the model's feature extraction ability for small targets, thereby obtaining more tomato feature information. Multiple sets of data are tested, and it is verified that the improved model can effectively solve missed or false detection of tomatoes in agricultural orchard environments. This test shows that the experiment achieves the research purpose of improving model detection accuracy.

The overall contribution of this paper is to combine those techniques in ripe tomato detection and prove the feasibility of this type of approach, providing technical support for later research on robotic arm grabbing tomatoes. It is of great significance to realize the development of agricultural automation and fruit grabbing.

II. RELATED WORK

Many scholars have conducted much research on the problem of small target detection. Their research methods have improved the accuracy of small target detection, but some problems are worth solving. Zhang et al. [12] proposed a multi-scale remote sensing small target detection method based on cosSTR-YOLOv7 to solve the problem of low target detection accuracy caused by too little feature information of small targets in geospatial remote sensing images. Constructing a new feature fusion layer in Neck reduces the loss of feature information. Adding a small target prediction layer in the prediction part improves the model's ability to detect small targets. The Map of the improved model is improved by 3.73%. However, when the target features are not obvious, there are still certain false or missed detections. Wang et al. [13] proposed a small target detection method based on improved YOLOv3-Tiny to solve difficult detection problems and low detection accuracy of small ship targets in remote sensing images. Mosaic data enhances and enriches the feature information of the target, and using the CBAM attention mechanism in the feature extraction part increases the feature extraction ability of the target. Experiments show that the improved model improves the accuracy. However, the problem of mutual occlusion between small targets in complex environments such as clouds and fog are caused, leading to the model's unsatisfactory detection effect. In order to solve the problems of object size change and complex background interference in the UAV aerial photography scene of the general target detection model, Ai et.al [14] added a neighborhood attention transformer in the last layer of the feature extraction network to retain global context information to extract more features. The CA attention mechanism is added to Neck to obtain channel and position information. Experiments show that the improved NATCA-Greater YOLO model map has improved by 2.9%. However, the detection effect is not good for objects with similar categories. In order to solve the problem that the infrared small target detection algorithm has noise influence in multiple scenes and the features are not obvious, Ni [15] used co-occurrence filtering as a trainable convolutional layer, and then designed it as a co-occurrence residual block in combination with residual ideas to improve the recognizability of infrared images during training. Whereas, due to the large amount of calculation of the model, the model's training process is affected. Li and Liu [16] conducted data enhancement based on open-source datasets to solve the problems of YOLOv5's poor detection of small targets. The channel attention ECA module is introduced to improve the recognition ability of the model. In order to improve the ability of the small target detection model, a small target detection layer is added to the Neck of the YOLOv5 model. The Map@.5 of the improved YOLOv5 model has been improved by 2.6%.

Xiao et al. [17] proposed an improved lightweight YOLOv3 target detection model for obstacle detection in the mine environment to reduce the occurrence of mine cart collision accidents. The residual network structure is added to the lightweight YOLOv3 model. Experiments show that the improved lightweight YOLOv3 model improves target detection accuracy. Wu et al. [18] replaced the backbone of YOLOv3 with Densenet for feature extraction. The improved YOLOv3 model effectively solved the problems of multi-target and multi-target occlusion. Detection accuracy was improved by 2.44%. Wang et al. [19] used Complete_IOU to replace the IOU in the traditional YOLOv3 to improve the recall and solved the problem of positive and negative samples by improving the confidence loss function. Experiments show that the improved YOLOv3 algorithm has better detection accuracy, an increase of 3.75%. Zhang et al. [20] used the INCEPTION module to process the deep features of the network to activate the multi-scale perception field and the Map of the improved YOLOv3 was 81%. Liu et al. [21] connected two ResNets to Resblock in the feature extraction network and optimized the network structure

darknet-53 by adding convolutional layers. Liu proposed an optimization training for UAV observation datasets for UAVs Methods. Experiments show that UAV-YOLO improves the detection accuracy by 0.33% than YOLOv3. Du et al. [22] proposed an ERF-YOLO algorithm to solve the problem that the detection effect of small targets cannot reach the expected effect. The ERF-YOLO algorithm proposes expanding the receptive field block to increase the range of information acquisition and upsampling the high-level information through deconvolution to obtain more feature information. The experiments show that the improvement of the latter YOLO has 4% higher Map than the original YOLOv2. Xianbao et al. [23] replaced the two-step down-sampling convolutional network in the original network structure with an image bi-segmented bi-linear up-sampling network to expand the eigenvalues. Adding a size recognition module to the input layer reduces the loss of morpheme features caused by no feature value filling. The residual network element is added to the output network layer to enhance the feature channel of small target detection. The experimental results show that the P is increased by 6.1%, and the Map is increased by 1.8%. Lim et al. [24] extracted the multi-scale features of the network's final output by splicing features and adding an attention mechanism. The experiments show that the Map reaches 78.1% on the PASCAL VOC2007 dataset. Ku et al. [25] added an ISR module to the network output layer to improve the resolution of the input image and replaced the remaining blocks in the backbone network with dense blocks to reduce the network structure parameters. Ku not only combined SPPnet and PANnet, but also added foreground to the loss function part and the background balance loss function. The experimental results show that the AP of the improved YOLO model is increased by 7.8%.

Table 1 summarizes the current status of previous research.

To sum up, previous research has improved the detection accuracy of the model to a certain extent, but there are still problems. For example, targets may not be detected in complex environments. The image features in the dataset are single and cannot meet the target detection requirements in the actual process. In addition, the model's generalization ability is insufficient. If the detection object is changed, the results may not be ideal. Therefore, we will start research from the model's structure and dataset to solve these problem. This paper uses Opencv-python code to expand the dataset by simulating the natural state of multi-feature targets under normal circumstances so that the model learns more target feature information. By optimizing the YOLO network structure, the feature extraction capability and detection accuracy of the model are improved. In addition, the open-source dataset called VOC2007 will be used to verify the generalization problem of the model.

YOLO is an end-to-end object detection method [26]. The YOLO series includes YOLOv1-YOLOv8 [27], but versions prior to YOLOv5 are too old to meet the experimental

TABLE 1. Research status.

Work	Dataset	Drawbacks
Build a new feature fusion layer and add a small target prediction layer [12]	Remote sensing dataset called DIOR	Targets with less obvious characteristics still have certain missed detections and misdetections.
Mosaic Data enhancement, CBAM[13]	Ship dataset	Targets affected by meteorological conditions such as clouds and fog cannot be accurately detected
Add neighborhood attention transformer, coordinate attention. Use Meta-ACON activation function[14]	Dataset called VisDrone2019-DET	For targets with similar categories, the detection effect is not good.
Use Co-occurrence Filter and Introduce a reverse attention mechanism[15]	Homemade infrared dataset	Due to the large amount of calculation of the model, the model's training process is affected.
Add Efficient Channel Attention and a small object detection layer. [16]	Homemade dataset	The evaluation indicators of the model are not complete enough.
Add residual network[17]	Mining Truck dataset	The detection accuracy of the model is not as good as the detection accuracy of the large model YOLOv3.
Replaced the backbone of YOLOv3 with Densenet[18]	The dataset is composed of pictures taken by the Internet and a small number of pictures taken by themselves.	The generalization ability of the model is insufficient.
Replace IOU in YOLOv3 with Complete_IOU and improve confidence loss function[19]	Homemade ship dataset	The model has too few evaluation indicators.
Use the inception module and optimize the loss function[20]	Homemade dataset	The dataset is too single.
Connect two ResNets to Resblock and optimize darknet-53 by adding convolutional layers[21].	Open-source dataset	The characteristics of the detected target are too single.
Expand the receptive field block and use deconvolution to obtain more feature information[22]	VOC2007 and VOC2012	The detected target has a single feature
Use an image bi-segmented bi-linear up-sampling network,add a size recognition module to the input layer and the residual network element[23]	VEDAI and DLR 3 K Munich vehicle.	Network depth is increased.

TABLE 1. *(Continued.)* **Research status.**

Extracte the multi-scale features of the network's final output by splicing features and adding an attention mechanism[24].	VOC2007	The detected target has a single feature.
Add an ISR module and replace the remaining blocks in the backbone network with dense blocks, etc. [25]	Self-constructed datasets	There is still a problem of insufficient feature extraction for difficult-to-detect samples or missing and false-positive cases.

requirements. Before selecting the experimental model, this paper conducts benchmark testing. The self-made dataset is used to conduct experimental comparisons of YOLOv5, YOLOv6, YOLOv7 and YOLOv8 models. According to experimental results, Regardless of params, GFLOPs or Map@.5, the performance of YOLOv5 and YOLOv6 series is not as good as YOLOv7. In the YOLOv8 series, although YOLOv8l and YOLOv8m can also achieve better Map@.5, they bring higher GFLOPs. Compared with other models, YOLOv7 can achieve high accuracy and achieve a balance between Params and GFLOPs.YOLOv7 has good performance So that YOLOv7 is selected as the experimental model to solve the occlusion problem in tomato detection.

YOLOv7 balances speed and accuracy perfectly, which makes itself favored by the industry [28]. It has faster detection speed and higher accuracy [29]. YOLOv7 consists of an input, backbone and head. The image is preprocessed at the input end and input to Backbone for feature extraction. The backbone layer comprises CBS, MP1, and ELAN [30]. CBS includes Conv, BN and SiLU. MP1 includes Maxpool and CBS. ELAN includes multiple CBS. The part of the head layer includes SPPCPC, Conv, ELAN, MPConv, and REP. SPPCSPC module and ELAN modules implement the feature extraction function [31]. The Head layer outputs feature maps of different sizes on the 75th, 88th, and 101st layers, and outputs prediction results through the reparameterized structure called REP layer.

The innovation of YOLOv7 lies in the use of module reparameterization and dynamic label assignment strategy, which has achieved high speed and accuracy. It achieves a balance between accuracy and reasoning performance. However, YOLOv7 adopts a new ELAN structure, MP structure and Silu activation function in the network structure. YOLOv7 proposes the E-ELAN structure [32] based on the ELAN structure, which realizes the continuous increase of the network learning ability without destroying the original gradient path. The network can learn more features and has stronger robustness. The MP module has two branches, which are used for downsampling. The first one goes through a maxpool, which is the maximum pooling. The function of maximum pooling is downsampling, and then a 1×1 convolution is performed to change the number of channels. The second one passes through a 1×1 convolution to change the number of channels, and then passes through a convolution block with a 3×3 convolution kernel and a step size of 2. This convolution block is used for downsampling. Finally, the results of the first branch and the second branch are added together to improve the feature extraction ability of the network. In addition, the dynamic label refers to the compound scaling of the YOLOv7 model. The new soft label method matches the detection frame with the prediction frame one by one.

YOLOV7 proposes a training method for the auxiliary head. The main purpose is to increase the training cost and improve accuracy without affecting the reasoning time, because the auxiliary head will only appear during the training process. YOLOv7 is mainly divided into two versions with and without auxiliary training heads. In the Python code, train.py is used for model training with the model including the auxiliary training head, and train_aux.py is used for model training without the auxiliary training head in the model.

III. IMPROVED MODEL

A. TOMATO-YOLOV7 MODEL

YOLOv7 is derived from the YOLOv4 and other model architectures [33]. In order to solve the problems of small tomatoes with few effective pixels, small scale, and lack of feature expression ability, the network structure of YOLOv7 is optimized. The improved YOLOv7 is named Tomato-YOLOv7, which is specially used for detecting tomatoes in agricultural orchards. As shown in Figure 1, it is the structure of the Tomato-YOLOv7. Because of the problem that the characteristics of the Occlusion are not obvious, methods for optimizing the network structure of YOLOv7 can improve the detection accuracy of the network model for tomatoes, and effectively reduce the false detection rate. The new model has made significant progress. It reduces the false detection rate of green plants around tomatoes and increases the feature extraction of tomatoes ability.

B. THE INNOVATION OF TOMATO-YOLOV7

To improve the Map@.5 of the model, the structure previously known as RepLKNet [34] undergoes a redesign and has been rebranded as ReplkDext. ReplkDext is used in the model, as shown in Figure 2(a), which indicates that ReplkDext is introduced into the backbone of YOLOv7. ReplkDext owns more shape information than the traditional CNN so that the model achieves a higher Map. Compared with the conventional 3×3 CNN, it has a larger kernel and obtains a larger effective receptive field and higher shape deviation. It can improve CNN feature extraction capabilities.

In order to solve the problem of the running speed of the network model, this paper redesigned the ELANN structure of the head layer of YOLOv7. Using FasterNet [35] (The location is shown in Figure 3) optimizes the structure between the traditional Concat and CBS. FasterNet has the advantage of reducing redundant computation and memory access to

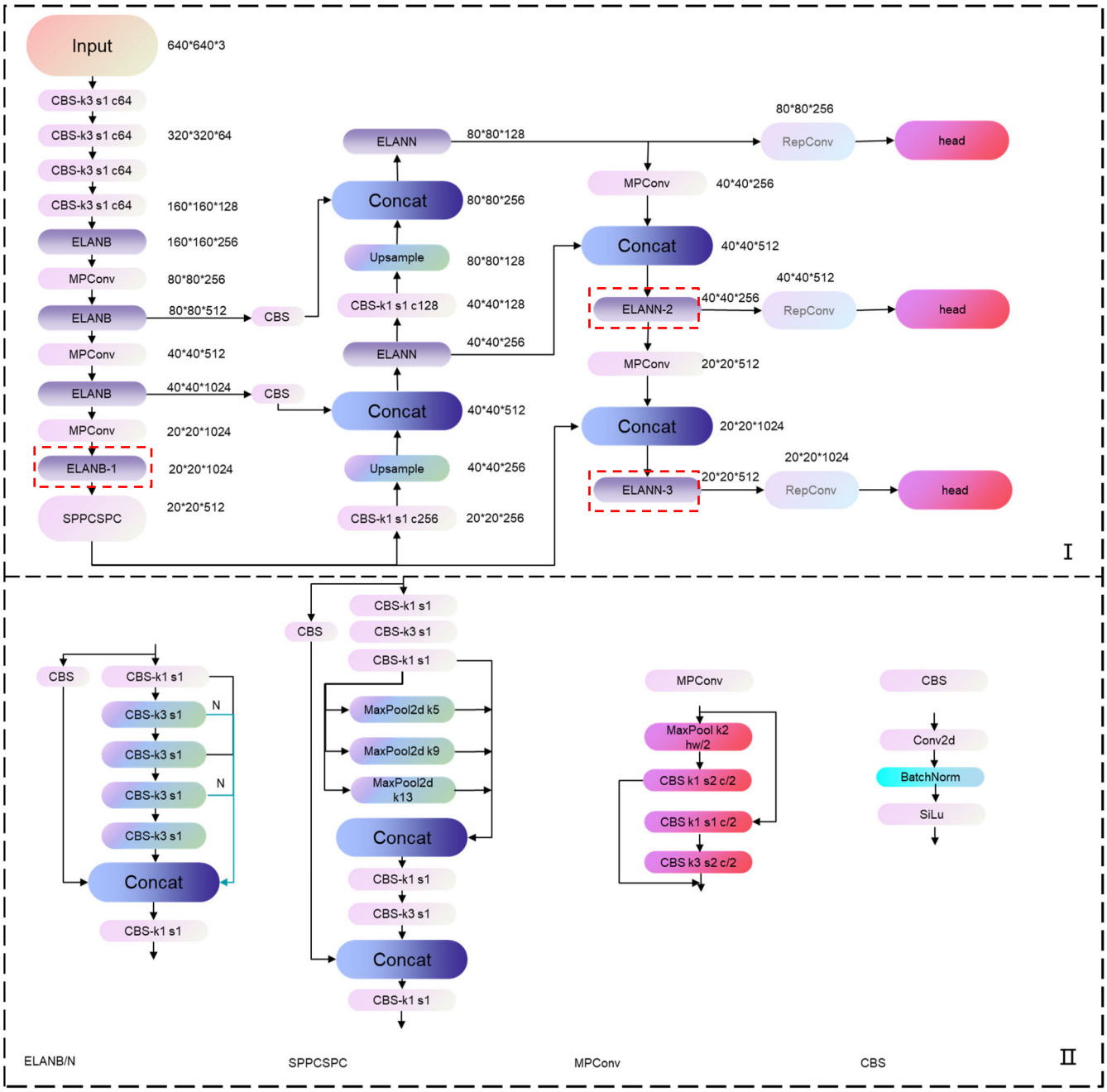

FIGURE 1. Structural of tomato-YOLOv7.

improve spatial feature extraction capabilities. The structure of FasterNet is shown in Figure 4. FasterNet overcomes the problem of low FLOPS caused by frequent access to memory in traditional neural networks, and can achieve a balance between running speed and detection accuracy. In view of the fairness of the experiment and to ensure the consistency of the experimental training parameters, the optimization of the 4 ELANN structures in the head layer is changed to the optimization of the last 2 ELANN structures, which can ensure the detection accuracy while improving the running speed. In Figure 4, Global Pook, Conv1 $\times$ 1 and FC are used for feature conversion and classification. The normalization layer (BN) and activation layer (ReLU) are placed between two Conv 1×1 to maintain feature diversity and achieve lower Delay. Compared with the base YOLOv7 model, the FPS of the Toma-to-YOLOv7 model has increased by 25.

To improve the ability of convolution, ODConv [36] is added after the last layer of ELANN-2 structure in the YOLOv7 network structure Head to improve the feature ex-traction ability of small targets and obtain more feature information of tomatoes. As shown in Figure 5(a), The improved ELANN-2 structure is named ELANN-3. As shown

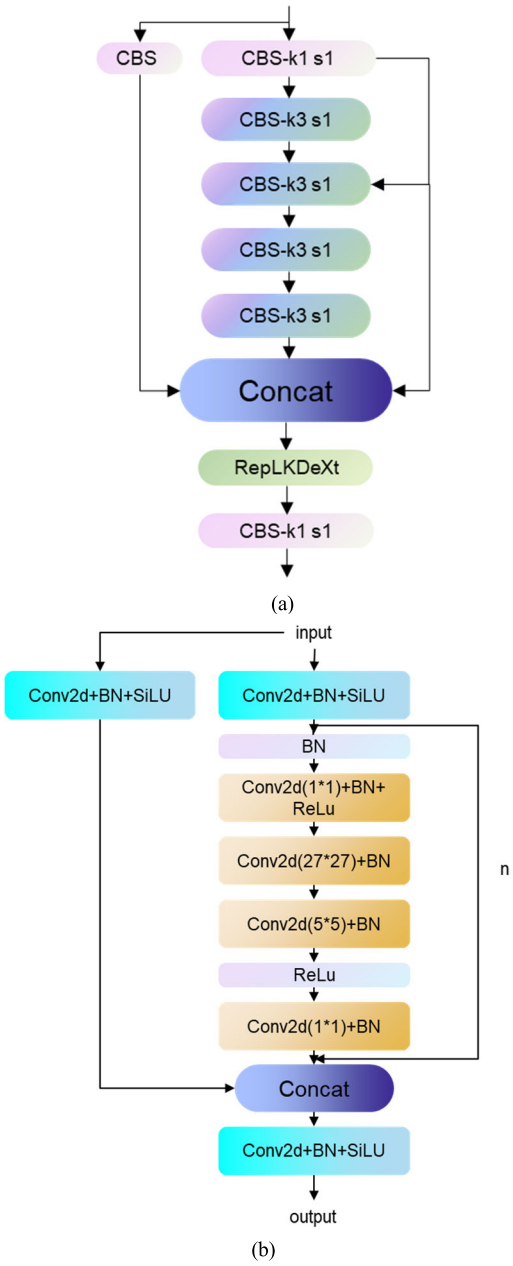

FIGURE 2. ReplkDext location information and structure of ReplkDext.

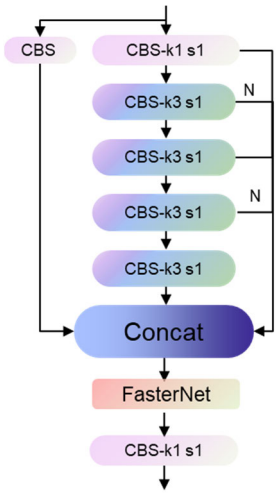

FIGURE 3. Structure location information of FasterNet.

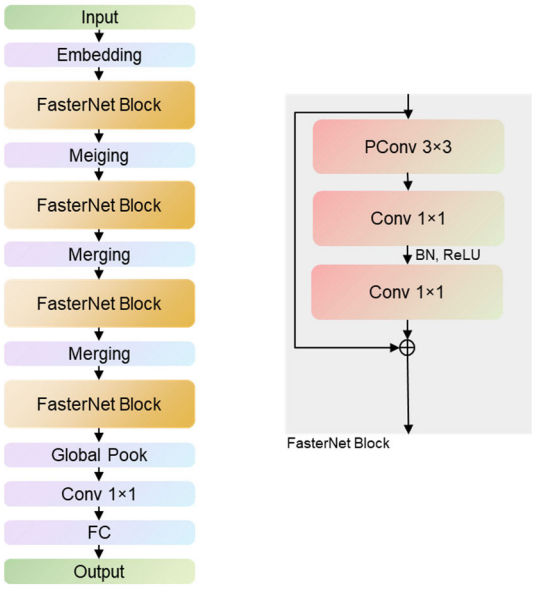

FIGURE 4. An illustration of the architecture and design of the FasterNet model structure.

in Figure5(b), the ODConv structure has the advantage that it can significantly enhance the feature extraction ability of CNN's basic convolution operation. ODConv utilizes a multi-dimensional attention mechanism to learn four types of attention for convolution kernels along all four dimensions of the kernel space in a parallel manner, and These attentions are gradually applied to the corresponding convolution kernels to improve performance.

IV. MATERIALS AND METHODS

A. TOMATO DATASET

Tomatoes are classified according to the national standard GH/T1193-2021.This paper collects ripe tomatoes in the natural environment of agricultural orchards by cameras [37]. Due to the presence of green plants around, the feature information of the target tomato is not fully displayed. The original dataset is expanded by Opencv on some of the images. Methods such as cropping pictures and stitching pictures can effectively simulate the state of tomatoes in the natural environment so that model gets more information. In addition, the Gaussian filter can effectively reduce the interference of camera noise to the experiment during the shooting process.

However, labelImg software is used to label the dataset's images and generate XML files. As shown in Figure 6, it is the labeling process of the dataset. The data set has 2410 images and 2410 XML files, which are divided into the training sets and verification sets according to the ratio of 8:2. Moreover the relevant configuration of the experimental environment is shown in Table 2.

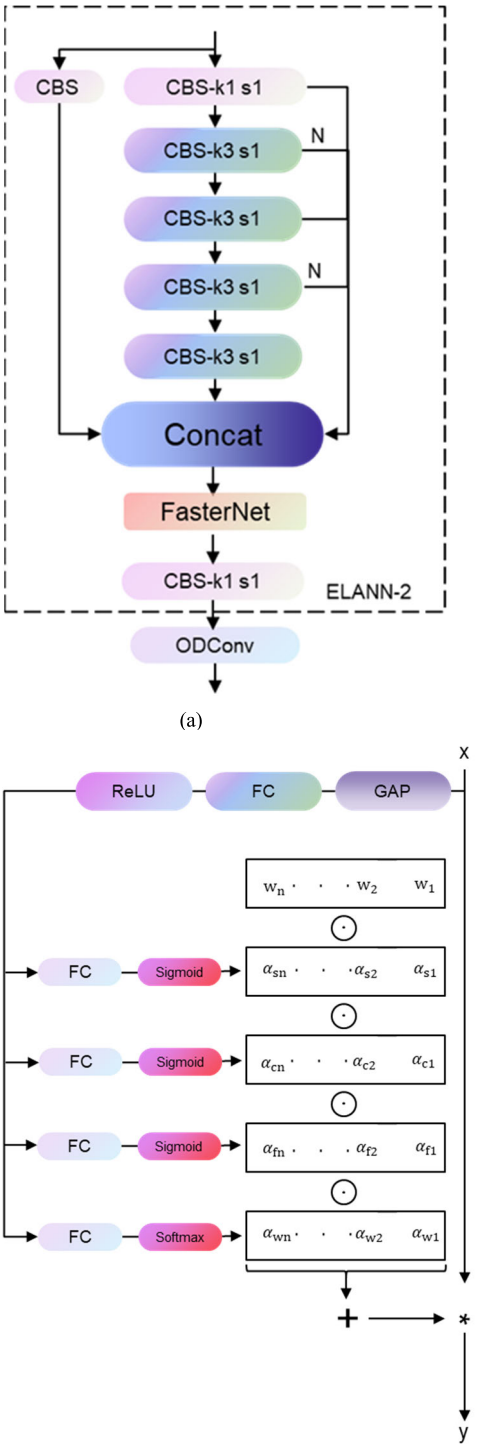

FIGURE 5. Location information and structure of ODConv.

B. MODEL TRAINING

The dataset is divided in proportion. The training set is 1913, and the verification set is 497. The training parameters of the model are as follows. The batch- size of the model is 32. Epochs is 200, and other parameters are the default parameters of the model. The default parameters include weights,

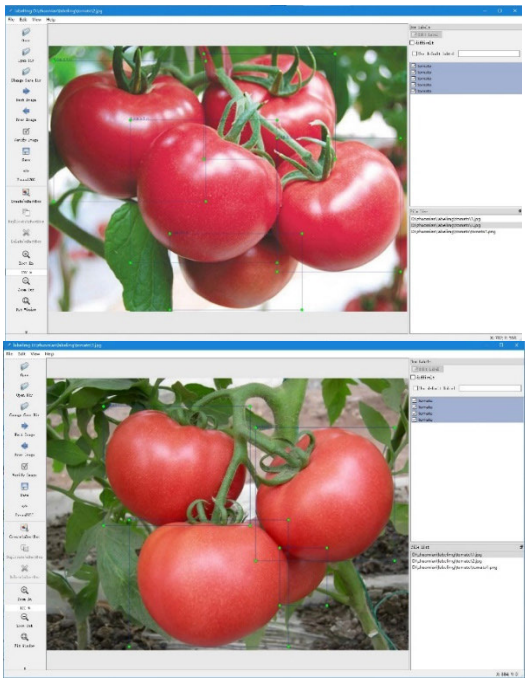

FIGURE 6. The labeling process of the tomato dataset.

TABLE 2. Experimental configuration of YOLOv7.

Software and hardware	Version
Pytorch	1.11.0
Python	3.8
Cuda	11.3
GPU	RTX 3090

cfg, data, img-size and others. Model training does not load pre-trained weights. Cfg represents the trained model. Data contains the relevant information of dataset. Img-size is set to 640 × 640. This paper utilizes various models for training. Mean average precision (Map)plays an important role in the model [38]. In order to verify the effectiveness of the improved model, an ablation experiment is carried out on the improved model.

C. EXPERIMENTAL COMPARISON

1) IMPROVEMENT EXPERIMENT OF YOLOV7

In this paper, aiming at the problem of inconspicuous features caused by the occlusion of tomatoes in agricultural orchards, the network structure of the YOLOv7 model is improved. The indicators of the improved model are shown in Table 3, and the table contains ablation experiments for the improved model.

2) COMPARATIVE EXPERIMENTS OF THE SAME DATASET AND DIFFERENT MODELS

To verify the effectiveness of the improved model, the Tomato-YOLOv7 model is compared with YOLOv5, YOLOv6 and YOLOv8. YOLOv5 has different versions to

TABLE 3. Experimental results of ablation experiments about YOLOv7 on the tomato dataset.

Model	FasterNet	ODConv	ReplkDext	P	R	Map@.5	Map@.5:0.95	FPS	Params(M)	GFLOPs(M)	Time (s)
YOLOv7				0.809	0.822	0.880	0.558	45	37.19	6.44	0.0099
YOLOv7-F	✓			0.813	0.846	0.889(+0.9%)	0.569	70	38.20	6.50	0.0104
YOLOv7-O		✓		0.819	0.839	0.888(+0.8%)	0.564	71	37.25	6.42	0.0095
YOLOv7-R			✓	0.801	0.850	0.887(+0.7%)	0.566	60	47.16	6.96	0.0112
YOLOv7-FO	✓	✓		0.799	0.848	0.891(+1.1%)	0.566	57	38.25	6.50	0.0111
YOLOv7-FR	✓		✓	0.808	0.832	0.885(+0.5%)	0.558	51	48.17	7.04	0.0122
YOLOv7-OR		✓	✓	0.831	0.825	0.889(+0.9%)	0.562	52	47.21	6.96	0.0121
Tomato-YOLOv7	✓	✓	✓	0.828	0.843	0.893(+1.3%)	0.568	42	48.22	7.04	0.0127

TABLE 4. Comparative analysis of various performance of different deep learning models.

Model	P	R	Map@.5	Map@.5:0.95	FPS	Params(M)	GFLOPs(M)	Time(s)
Tomato-YOLOv7	0.828	0.843	0.893	0.568	42	48.22	7.04	0.0127
YOLOv5l	0.821	0.826	0.880	0.575	75	46.14	6.63	0.0158
YOLOv5m	0.807	0.802	0.864	0.560	97	20.87	2.95	0.0129
YOLOv5s	0.774	0.793	0.837	0.530	126	7.02	0.98	0.0098
YOLOv5n	0.774	0.779	0.844	0.522	138	1.77	0.26	0.0091
YOLOv6l	0.702	0.702	0.768	0.494	47	110.86	391.2	0.0184
YOLOv6m	0.721	0.760	0.795	0.512	94	51.98	161.1	0.0165
YOLOv6s	0.836	0.789	0.871	0.567	133	16.30	44.00	0.0114
YOLOv6n	0.761	0.821	0.855	0.551	135	4.23	11.80	0.0111
YOLOv8l	0.837	0.820	0.879	0.593	76	43.61	164.8	0.0177
YOLOv8m	0.854	0.800	0.879	0.586	94	25.84	78.70	0.0158
YOLOv8s	0.817	0.801	0.871	0.568	120	11.13	28.40	0.0134
YOLOv8n	0.783	0.800	0.854	0.553	129	3.00	8.10	0.0115

adapt to different computing power and real-time requirements. Yolov5l, yolov5m, yolov5s and yolov5n only differ in model depth (number of C3 modules) and width (number of channels in the network), and everything else is the same. The ''l'' in YOLOv5l stands for ''large'', and this model is used on devices with strong computing capabilities. The ''m'' of YOLOv5m stands for ''medium'', and this model is used on devices with certain computing capabilities. The ''s'' of YOLOv5s stands for ''small'', and this model is used on mobile devices or edge devices. YOLOv5n is exclusively used on Nano devices. The same is true for YOLOv6 and YOLOv8.

As shown in Table 4, it is obvious that the Map@.5 of the Tomato-YOLOv7 is 89.3, which is higher than that of other models. Under the premise of ensuring the detection accuracy of Tomato, the Params, and GFLOPs of the Tomato-YOLOv7 are superior to other YOLO models. It verifies the effectiveness of the improvement.

Whether a model is more effective than other contemporary deep learning methods is mainly judged through some evaluation indicators. Among them, Map@.5 is the main indicator for evaluating the YOLO model. As shown in Table 4, Although YOLOv5 has similar Params and GFLOPs, the main evaluation index Map@.5 is lower than Tomato-YOLOv7's Map@.5. Therefore, the performance of Tomato-YOLOv7 is better than YOLOv5. Regardless of Map@.5, Params and GFLOPs, the performance of YOLOv6 is not as good as Tomato-YOLOv7. Map@.5, Params and GFLOPs of Tomato-YOLOv7 can outperform other models. It is obvious from Table 4 that while the Map@.5 of YOLOv8 is improved, the GFLOPs quickly increase to 164.8. This result is not good. In comparison, Tomato-YOLOv7 has better performance.

3) GENERALIZATION OF TOMATO-YOLOV7 MODEL

The generalization of the model is an important indicator for evaluating the mod-el. The generalization of the model is judged according to the degree to which the same model uses different datasets to improve the performance of the model. This paper verifies the model's generalization using the open-source dataset called VOCtest_06-Nov-2007. As shown in Table 5, the P and Map of the Tomato-YOLOv7 are better than the YOLOv7, and the experimental results show the model's generalization. The improved model in this paper is suitable for applications in related fields.

4) DETECTION RESULTS OF DIFFERENT MODELS

Figure 7 shows the detection results of different models. Images on the left are the result of YOLOv7. Images on the right are the result of Tomato-YOLOv7. The dataset collects ripe tomatoes, so the green tomatoes in the picture cannot be detected. There are missed detections in the image on the left. On the contrary, the right image has a better detection effect. It is obvious that the improved model in this paper is effective.

TABLE 5. Generalization experiments of the model.

Dataset	Models	P	R	Map@.5	Map@.5:0.95
VOCtest_06-Nov-2007	YOLOv7	0.742	0.671	0.730	0.48
	Tomato-YOLOv7	0.787	0.669	0.734	0.501

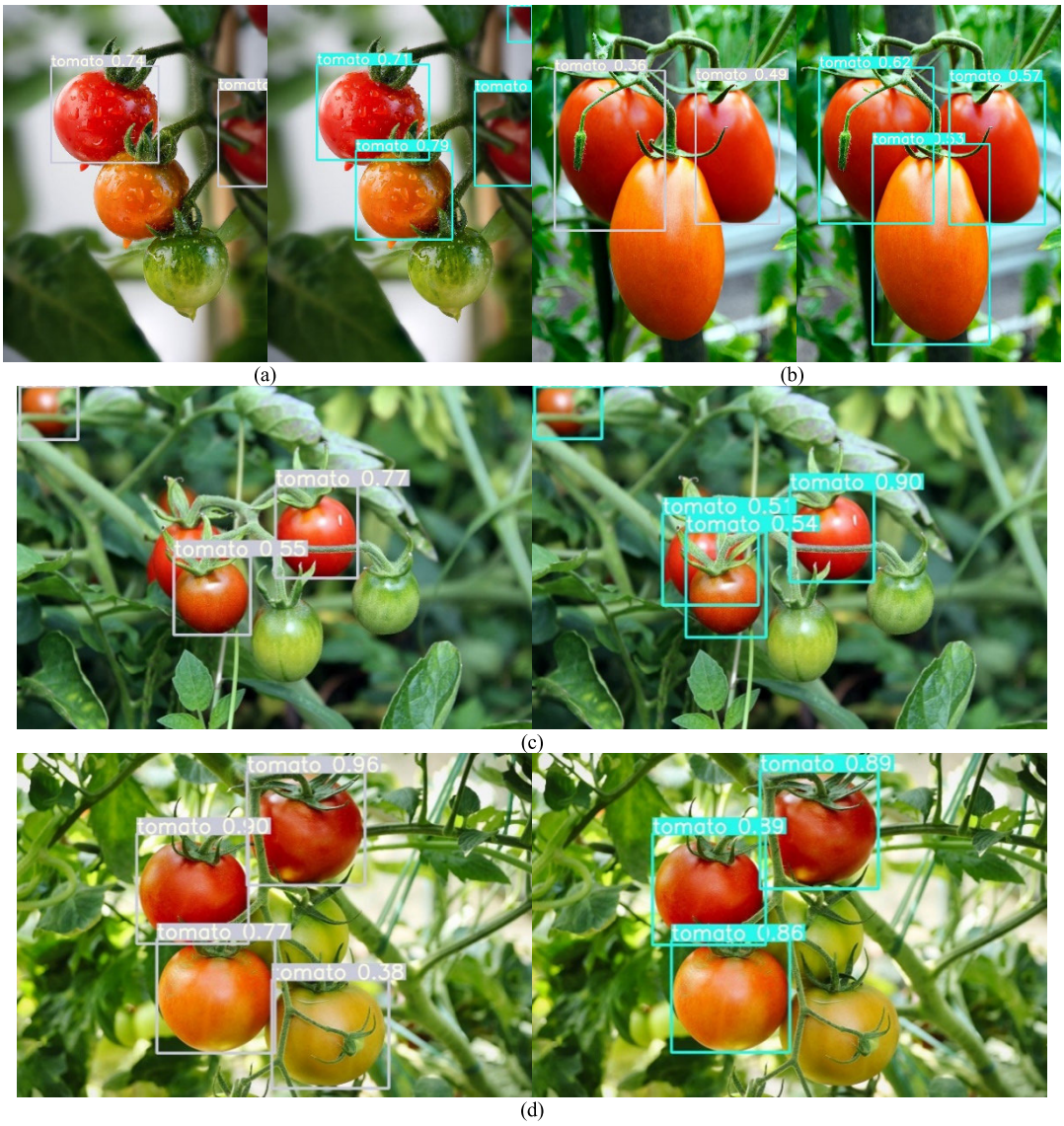

FIGURE 7. Image detection through different models.

V. RESULTS AND DISCUSSION

A. COMPARISON OF DIFFERENT MODELS ON MAP

The epochs of the model proposed in the experiment is 300. Since the training results show that the model tends to be stable when the epochs are around 200. So epochs of training are 200.The map is an important indicator of model evaluation during the training, and its meaning is the average detection accuracy of the model. The Map formula is shown as follows:

$$Map = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} AP_i \tag{1}$$

As shown in Figure(a), they are the Map of YOLOv7 and Tomato-YOLOv7 during the training. Whether it is Map@.5 or Map@.5: 0.95, the Tomato-YOLOv7 is higher than the

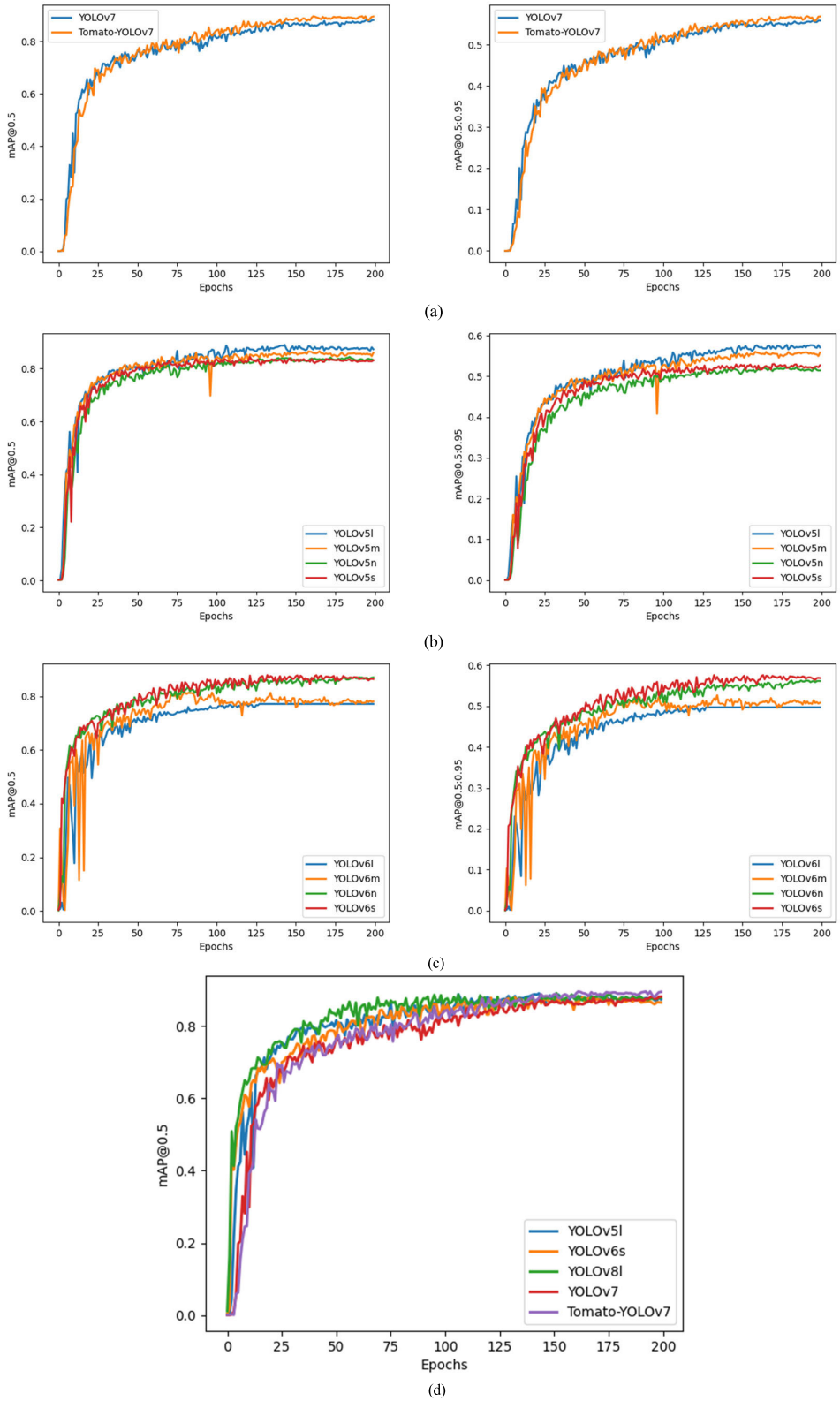

FIGURE 8. Comparison of different maps curves of various models.

basic model YOLOv7.So it verifies the effectiveness of the model improvement. As shown in Figure 8(b) and 8(c), It is the comparison of other models on Map.

As shown in Figure(d), it compares Tomato-YOLOv7, YOLOv7, YOLOv5l, YOLOv6s, and YOLOv8l. It is clear that the Tomato-YOLOv7 model outperforms other models.

In order to better improve the detection accuracy, the model controls the stability of the model during training, which promotes the slow growth of the model's Map@.5. After Epochs>=125, the growth rate of model accuracy increases and is much higher than other models. Tomato-YOLOv7 adds about 11M params and about 0.6 GFLOPs, in exchange for a 1.3% accuracy improvement.

B. COMPARISON OF DIFFERENT MODELS ON FPS

In target detection, FPS is an important indicator for real-time detection. For example, 50FPS means there are 50 frames per second. From the Table 3, the FPS of YOLOv7 is 45 while the FPS of Tomato-YOLOv7 is 42. It meets the requirements of real-time detection. From Table 4, it can be seen that although other YOLO models have achieved good FPS, but ignoring Map. the Map of the model is not as good as that of the Tomato-YOLOv7 model.

C. COMPARISON OF DIFFERENT MODELS ON PARAMS AND GFLOPS

In Table 3, the parameters of YOLOv7 and related models in the ablation experiment are stable at 37 to 48, and the GFLOPs are stable at 6.44 to 7.04. The Tomato-YOLOv7 increases the detection accuracy of Tomato by 1.3% based on adding a small amount of param and GFLOPs.

In the comparative analysis of Tomato-YOLOv7 and YOLOv5, YOLOv6, and YOLOv8, the performance of Tomato-YOLOv7 has an advantage about Map@.5, params, and GFLOPs. Because YOLO experimental research mainly pursues Map. It is evident from Table 4 that although YOLOv5m, YOLOv5s and YOLOv5n have very low params and GFLOPs, the main Map@.5 is very bad and cannot reach the expected value. The YOLOv6 series fail to achieve a good balance on Map@.5, params and GFLOPs in this paper. In addition, YOLOv8s and YOLOv8n also have very low params and GFLOPs, but the main Map@.5 is not ideal.it is far from the performance of Tomato-YOLOv7. No matter in terms of Map@.5, params and GFLOPs, the performance of the Tomato-YOLOv7 is obviously superior to other models. Map@.5 is much higher than other models by at least 1.3%, and params and GFLOPs can reach a balance. This reflects the effectiveness of the model improvement.

D. TIME CONSUMED BY THE MODEL

From Table 3, it is obvious to draw a conclusion that the improved Tomato-YOLOv7 model has a similar time to YOLOv7. The entire time is 0.01 seconds. and the time floats on this basis. By rounding the times in Table 4, it is not difficult to find that the time for YOLOv5l, YOLOv6l, YOLOv6m, YOLOv8l and YOLOv8m is 0.02 seconds, and the other models are all 0.01 seconds. Compared with other models, the detection accuracy is improved and Time will increase at the same time. The model proposed in this paper achieves a runtime of 0.01 seconds while maintaining higher accuracy. Therefore, the improved model has more advantages.

E. GENERALIZATION OF THE MODEL

In this paper, the experiment on YOLOv7 is carried out on the self-made dataset Tomato, and the result shows that the detection accuracy of the Tomato-YOLOv7 is improved. However, the generalization of the model is also an important indicator for evaluating a model. The model's generalization means that other datasets can also improve the detection accuracy of the Tomato-YOLOv7. The dataset is replaced with the public dataset called VOCtest_06-Nov-2007. Previous experiment is repeated. According to Table 5, the Map@.5:0.95 of the Tomato-YOLOv7 has been steadily improved by 2.1%, and Tomato-YOLOv7 has good generalization. Therefore, it can be extended to other Target Detection and has good theoretical significance.

F. LIMITATION OF STUDY

The limitation of this study is the number of images about tomatoes in different environments. For example, tomatoes grow in rainy, foggy, snowy and other complex weather conditions. Due to limited experimental conditions, the images in the dataset were taken in a single environment. Next study will expand the number of images in the dataset. Taking more pictures in different complex environments is used to increase the diversity of tomatoes under study and enhance the generality of the improved model.

VI. CONCLUSION

In order to solve the problem of missed detection or false detection in ripe tomato detection, this paper proposes a model based on improved YOLOv7. This paper creates a dataset about ripe tomatoes, and extracts some images from the dataset for some image processing. Image processing includes Gaussian filtering that reduces the image's noise and suppresses the external environment's influence on the experimental results. The model has been improved as follows. ReplkDext is added to the backbone layer of the YOLOv7, which obtains more shape information and enhances feature extraction capabilities of model. It improves Map@.5 by 0.7%. Using FasterNet optimizes the structure of ELANN in the Head layer, which reduces redundant calculation and memory access to enhance the ability of spatial feature extraction. The model achieves a balance between running speed and detection accuracy. It improves Map@.5 by 0.9%. ODConv uses a multi-dimensional attention mechanism and a parallel method to learn the four types of attention of the convolution kernel and apply it to the corresponding convolution kernel to improve the feature extraction ability of the convolution and operation. It can get more characteristic information about Tomato. It improves Map@.5 by 0.8%. The Map@.5 of Tomato-YOLOv7 is 1.3% higher than that of YOLOv7. The improved model can effectively solve the false or missed detection caused by occlusion in tomato detection. In addition, the FPS of Tomato-YOLOv7 reaches 42, which meets the requirements of real-time detection.

The scenes where tomatoes are located in the data set of this article are limited. Follow-up research can add tomato data sets in multiple environments (such as rainy days, snowy days and foggy days) to increase the quality of the data set. In future research, research can be conducted towards tomato target detection in more complex environments. Object detection in complex situations is the next research goal.

REFERENCES

[1] G. Yang, J. Wang, Z. Nie, H. Yang, and S. Yu, "A lightweight YOLOv8 tomato detection algorithm combining feature enhancement and attention," *Agronomy*, vol. 13, no. 7, p. 1824, Jul. 2023, doi: 10.3390/agronomy13071824.

[2] L. Jia, T. Wang, Y. Chen, Y. Zang, X. Li, H. Shi, and L. Gao, "MobileNet-CA-YOLO: An improved YOLOv7 based on the MobileNetV3 and attention mechanism for Rice pests and diseases detection," *Agriculture*, vol. 13, no. 7, p. 1285, Jun. 2023, doi: 10.3390/agriculture13071285.

[3] X. Xu, Z. Xue, and Y. Zhao, "Research on an algorithm of express parcel sorting based on deeper learning and multi-information recognition," *Sensors*, vol. 22, no. 17, p. 6705, Sep. 2022, doi: 10.3390/s22176705.

[4] G. Yin, M. Yu, M. Wang, Y. Hu, and Y. Zhang, "Research on highway vehicle detection based on faster R-CNN and domain adaptation," *Appl. Intell.*, vol. 52, pp. 3483–3498, Jul. 2021, doi: 10.1007/s10489-021-02552-7.

[5] J. Bai, "Research on the improvement of target detection algorithm based on YOLOv3," M.S. thesis, School Comput. Sci. Technol., North Univ. China, Shanxi Province, Taiyuan City, China, 2022.

[6] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, Y. Li, B. Zhang, Y. Liang, L. Zhou, X. Xu, X. Chu, X. Wei, and X. Wei, "YOLOv6: A single-stage object detection framework for industrial applications," 2022, *arXiv:2209.02976*.

[7] C.-Y. Wang, A. Bochkovskiy, and H.-Y. Mark Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.

[8] X. Guo, Y. Zhu, and S. Li, "Scale adaptive small target recognition algorithm in complex agricultural environment-taking bees as research objects," *Smart Agricult. Chin. English*, vol. 4, no. 1, pp. 140–149, Mar. 2022, doi: 10.12133/j.smartag.SA202203003.

[9] H. Yang, Y. Liu, S. Wang, H. Qu, N. Li, J. Wu, Y. Yan, H. Zhang, J. Wang, and J. Qiu, "Improved apple fruit target recognition method based on YOLOv7 model," *Agriculture*, vol. 13, no. 7, p. 1278, Jun. 2023, doi: 10.3390/agriculture13071278.

[10] F. Chen, L. Zhang, and S. Kang, "Soft-NMS-enabled YOLOv5 with SIOU for small water surface floater detection in UAV-captured images," *Sustainability*, vol. 15, no. 14, Jul. 2023, Art. no. 10751, doi: 10.3390/su151410751.

[11] L. Yuan, "Research on video small target detection algorithm based on deep learning," M.S. thesis, College Big Data Inf. Eng., Guizhou Univ., Guiyang City, Guizhou, China, 2021.

[12] X. Zhang, Z. Zhu, and Y. Guo, "Multi-scale remote sensing small target detection based on cosSTR-YOLOv7," *Electro-Opt. Control*, pp. 1–9, Jun. 2023. [Online]. Available: https://kns.cnki.net/kcms2/detail/41.1227.tn.20230615.1017.002.html

[13] X. Wang, T. Jiang, and Z. Ma, "Small remote sensing ship target detection method based on improved YOLOv3-tiny," *Comput. Times*, vol. 3, pp. 111–115, Jan. 2023, doi: 10.16644/j.cnki.cn33-1094/tp.2023.03.026.

[14] Z. Ai, S. Zang, and M. Chen, "Small target detection in aerial photography based on NATCA-Greater YOLO," *J. Qingdao Univ. Eng. Technol. Ed.*, vol. 38, no. 2, pp. 18–25, Jun. 2023, doi: 10.13306/j.1006-9798.2023.02.003.

[15] H. Ni, "Research on key algorithms for infrared small target detection based on deep learning," M.S. thesis, School Commun. Inf. Eng., Nanjing Univ. Posts Telecommun., Nanjing, Jiangsu, China, 2022.

[16] D. Li, and H. Liu, "YOLOv5 helmet detection algorithm for small targets," *Modern Inf. Technol.*, vol. 7, no. 9, pp. 9–13, May 2023, doi: 10.19850/j.cnki.2096-4706.2023.09.002.

[17] D. Xiao, F. Shan, Z. Li, B. T. Le, X. Liu, and X. Li, "A target detection model based on improved tiny-YOLOv3 under the environment of mining truck," *IEEE Access*, vol. 7, pp. 123757–123764, 2019, doi: 10.1109/ACCESS.2019.2928603.

[18] F. Wu, G. Jin, M. Gao, Z. HE, and Y. Yang, "Helmet detection based on improved YOLO v3 deep model," presented at the *Proc. IEEE 16th Int. Conf. Netw., Sens. Control (ICNSC)*, May 2019, pp. 363–368. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8743246

[19] F. Wang, X. Yang, Y. Zhang, and J. Yuan, "Ship target detection algorithm based on improved YOLOv3," in *Proc. 3rd Int. Conf. Big Data Technol.*, Sep. 2020, pp. 162–166, doi: 10.1145/3422713.3422721.

[20] B. Zhang, X. Qian, R. Yang, and Z. Xu, "Water surface target detection based on improved YOLOv3 in UAV images," in *Proc. 9th Int. Conf. Commun. Broadband Netw.*, Feb. 2021, pp. 47–53, doi: 10.1145/3456415.3456424.

[21] M. Liu, X. Wang, A. Zhou, X. Fu, Y. Ma, and C. Piao, "UAV-YOLO: Small object detection on unmanned aerial vehicle perspective," *Sensors*, vol. 20, no. 8, p. 2238, Apr. 2020, doi: 10.3390/s20082238.

[22] Z. Du, J. Yin, and J. Yang, "Expanding receptive field YOLO for small object detection," *J. Phys., Conf. Ser.*, vol. 1314, no. 1, Oct. 2019, Art. no. 012202, doi: 10.1088/1742-6596/1314/1/012202.

[23] C. Xianbao, Q. Guihua, J. Yu, and Z. Zhaomin, "An improved small object detection method based on YOLO v3," *Pattern Anal. Appl.*, vol. 24, no. 3, pp. 1347–1355, May 2021, doi: 10.1007/s10044-021-00989-7.

[24] J.-S. Lim, M. Astrid, H.-J. Yoon, and S.-I. Lee, "Small object detection using context and attention," presented at the *Proc. Int. Conf. Artif. Intell. Inf. Commun. (ICAIIC)*, Apr. 2021, pp. 181–186. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9415217

[25] B. Ku, K. Kim, and J. Jeong, "Real-time ISR-YOLOv4 based small object detection for safe shop floor in smart factories," *Electronics*, vol. 11, no. 15, p. 2348, Jul. 2022, doi: 10.3390/electronics11152348.

[26] J. Terven and D. Cordova-Esparza, "A comprehensive review of YOLO: From YOLOv1 and beyond," 2023, *arXiv:2304.00501*.

[27] M. Anwar, Y. Kristian, and E. Setyati, "Klasifikasi penyakit tanaman cabai rawit dilengkapi dengan segmentasi citra daun dan buah menggunakan YOLO v7," *INTECOMS: J. Inf. Technol. Comput. Sci.*, vol. 6, no. 1, pp. 540–548, Jun. 2023, doi: 10.31539/intecoms.v6i1.6071.

[28] Y. Wang, H. Wang, and Z. Xin, "Efficient detection model of steel strip surface defects based on YOLO-V7," *IEEE Access*, vol. 10, pp. 133936–133944, 2022, doi: 10.1109/ACCESS.2022.3230894.

[29] D. Wu, S. Jiang, E. Zhao, Y. Liu, H. Zhu, W. Wang, and R. Wang, "Detection of camellia oleifera fruit in complex scenes by using YOLOv7 and data augmentation," *Appl. Sci.*, vol. 12, no. 22, p. 11318, Nov. 2022, doi: 10.3390/app122211318.

[30] Y. Zhang, Y. Sun, Z. Wang, and Y. Jiang, "YOLOv7-RAR for urban vehicle detection," *Sensors*, vol. 23, no. 4, p. 1801, Feb. 2023, doi: 10.3390/s23041801.

[31] J. Chen, S. Bai, G. Wan, and Y. Li, "Research on YOLOv7-based defect detection method for automotive running lights," *Syst. Sci. Control Eng.*, vol. 11, no. 1, Mar. 2023, Art. no. 2185916, doi: 10.1080/21642583.2023.2185916.

[32] I. Gallo, A. U. Rehman, R. H. Dehkordi, N. Landro, R. La Grassa, and M. Boschetti, "Deep object detection of crop weeds: Performance of YOLOv7 on a real case dataset from UAV images," *Remote Sens.*, vol. 15, no. 2, p. 539, Jan. 2023, doi: 10.3390/rs15020539.

[33] M. J. A. Soeb, M. F. Jubayer, T. A. Tarin, M. R. Al Mamun, F. M. Ruhad, A. Parven, N. M. Mubarak, S. L. Karri, and I. M. Meftaul, "Tea leaf disease detection and identification based on YOLOv7 (YOLO-T)," *Sci. Rep.*, vol. 13, no. 1, p. 6078, Apr. 2023, doi: 10.1038/s41598-023-33270-4.

[34] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31×31: Revisiting large kernel design in CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11963–11975. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/html/Ding_Scaling_Up_Your_Kernels_to_31x31_Revisiting_Large_Kernel_Design_CVPR_2022_paper.html

[35] J. Chen, S.-H. Kao, H. He, W. Zhuo, S. Wen, C.-H. Lee, and S.-H.-G. Chan, "Run, don't walk: Chasing higher FLOPS for faster neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 12021–12031. https://openaccess.thecvf.com/content/CVPR2023/html/Chen_Run_Dont_Walk_Chasing_Higher_FLOPS_for_Faster_Neural_Networks_CVPR_2023_paper.html

[36] C. Li, A. Zhou, and A. Yao, "Omni-dimensional dynamic convolution," 2022, *arXiv:2209.07947*.

[37] Y. Li, J. Guo, X. Guo, K. Liu, W. Zhao, Y. Luo, and Z. Wang, "A novel target detection method of the unmanned surface vehicle under all-weather conditions with an improved YOLOV3," *Sensors*, vol. 20, no. 17, p. 4885, Aug. 2020, doi: 10.3390/s20174885.
[38] M. Durve, S. Orsini, A. Tiribocchi, A. Montessori, J.-M. Tucny, M. Lauricella, A. Camposeo, D. Pisignano, and S. Succi, "Benchmarking YOLOv5 and YOLOv7 models with DeepSORT for droplet tracking applications," *Eur. Phys. J. E*, vol. 46, no. 5, p. 32, May 2023, doi: 10.1140/epje/s10189-023-00290-x.

JUN GUO is currently an Assistant Professor with the School of Automobile Engineering, Yancheng Institute of Technology, China. He presides over multiple projects of enterprise. His current research interests include agricultural engineering and the mechanical properties of components.

YUE YANG is currently pursuing the master's degree with the School of Automobile Engineering, Yancheng Institute of Technology, China. His current research interests include agricultural machinery design and machine vision.

XINYAN LIN is currently a Teacher with the School of Automotive Engineering, Yancheng Institute of Technology, Yancheng, Jiangsu, China. He has extensive teaching experience in image processing. His current research interests include object detection and image processing.

MUHAMMAD SOHAIL MEMON is currently an Assistant Professor with the Faculty of Agricultural Engineering, Sindh Agriculture University, Pakistan. His current research interests include agricultural engineering, remote sensing/GIS applications, machine learning, and farm mechanization.

WEI LIU is currently the Dean of the School of Automotive Engineering, Yancheng Institute of Technology, Yancheng, Jiangsu, China. His current research interest includes vehicle detection.

MEIQI ZHANG is currently an Associate Professor with the School of Automotive Engineering, Yancheng Institute of Technology, Yancheng, Jiangsu, China. His current research interests include deep learning and new energy electronic control technology.

ENHUI SUN is currently pursuing the master's degree with the School of Automobile Engineering, Yancheng Institute of Technology, China. Her current research interests include machine vision and force research on agricultural vehicles.