Fast and accurate detection of kiwifruits in the natural environment using improved YOLOv4

Jinpeng Wang[1,2*], Lei Xu[1,2], Song Mei[3*], Haoruo Hu[1], Jialiang Zhou[1], Qing Chen[1]

(1. *Co-Innovation Center of Efficient Processing and Utilization of Forest Resources, Nanjing Forestry University, Nanjing 210037, China*;
2. *School of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, China*;
3. *Nanjing Institute of Agricultural Mechanization, Ministry of Agricultural and Rural Affairs, Nanjing, 210014, China*)

Abstract: Real-time detection of kiwifruits in natural environments is essential for automated kiwifruit harvesting. In this study, a lightweight convolutional neural network called the YOLOv4-GS algorithm was proposed for kiwifruit detection. The backbone network CSPDarknet-53 of YOLOv4 was replaced with GhostNet to improve accuracy and reduce network computation. To improve the detection accuracy of small targets, the upsampling of feature map fusion was performed for network layers 151 and 154, and the spatial pyramid pooling network was removed to reduce redundant computation. A total of 2766 kiwifruit images from different environments were used as the dataset for training and testing. The experiment results showed that the F1-score, average accuracy, and Intersection over Union (IoU) of YOLOv4-GS were 98.00%, 99.22%, and 88.92%, respectively. The average time taken to detect a 416×416 kiwifruit image was 11.95 ms, and the model's weight was 28.8 MB. The average detection time of GhostNet was 31.44 ms less than that of CSPDarknet-53. In addition, the model weight of GhostNet was 227.2 MB less than that of CSPDarknet-53. YOLOv4-GS improved the detection accuracy by 8.39% over Faster R-CNN and 8.36% over SSD-300. The detection speed of YOLOv4-GS was 11.3 times and 2.6 times higher than Faster R-CNN and SSD-300, respectively. In the indoor picking experiment and the orchard picking experiment, the average speed of the YOLOv4-GS processing video was 28.4 fps. The recognition accuracy was above 90%. The average time spent for recognition and positioning was 6.09 s, accounting for about 29.03% of the total picking time. The overall results showed that the YOLOv4-GS proposed in this study can be applied for kiwifruit detection in natural environments because it improves the detection speed without compromising detection accuracy.

Keywords: kiwifruits, fruit recognition, natural environments, YOLOv4
DOI: 10.25165/j.ijabe.20241705.7658

Citation: Wang J P, Xu L, Mei S, Hu H R, Zhou J L, Chen Q. Fast and accurate detection of kiwifruits in the natural environment using improved YOLOv4. Int J Agric & Biol Eng, 2024; 17(5): 222–230.

1 Introduction

Kiwifruit is an important cash crop worldwide, and the kiwifruit industry drives China's agricultural economy[1]. At present, kiwifruit is mainly picked by hand, which is time-consuming and labor-intensive[2,3]. For kiwifruit picking automation, the machine vision system is a key component of the picking robot[4]. The speed and accuracy of fruit recognition determine the efficiency and stability of the picking robot[5,6]. Therefore, the rapid recognition and accurate positioning of kiwifruits in the natural environment is of great significance for intelligent picking robots.

Currently, traditional image processing methods, such as the Hough transform[7], *K*-means clustering algorithm[8], and Sobel edge extraction system[9], have mostly been used for fruit recognition. These methods mostly use feature description methods to obtain the color, texture, shape, and other features of the fruit for complete recognition. However, kiwifruit images collected in natural environments are generally obtained under varying lighting conditions, and these kiwifruits are often occluded by branches and leaves. This leads to an increase in the false detection of kiwifruits in the natural environment, which hinders their effective extraction.

Convolutional neural networks can learn feature information very well, which gives the features better generalization, classification, and characterization recognition abilities[10-12]. At present, convolutional neural networks have been widely applied in target detection[13,14], fruit recognition, and hyperspectral analysis[15] of intelligent agriculture. The target detection methods based on convolutional neural networks are mainly 1) the two-stage methods represented by the Region-based Convolutional Neural Networks (R-CNN) series[16,17] and 2) the single-stage methods represented by the Single Shot Multibox Detector (SSD)[18] and You Only Look Once (YOLO)[19]. Xiong et al.[20] used the Faster R-CNN model for the visual detection of green citrus on trees; the average detection accuracy of their training model on the test set was 85.49%. Song et al.[21] used the Faster R-CNN model based on VGG16 to detect kiwifruits; the average detection time of each image was 347 ms, and the average detection accuracy was 87.61%. The two-stage detection method had better detection accuracy than the traditional recognition method, but the detection time was longer. Li et al.[22] used an improved SSD model for citrus recognition, and the improved recognition model achieved an average accuracy of

Received date: 2022-06-19 **Accepted date:** 2022-10-30
Biographies: Lei Xu, Master, research interest: agricultural engineering, Email: 570486211@qq.com; **Haoruo Hu**, Master, research interest: fruit picking robot design, Email: 1113949361@qq.com; **Jialiang Zhou**, Master, research interest: vision system for fruit picking robot, Email: zhoujialiang@njfu.edu.cn; **Qing Chen**, PhD, Associate Professor, research interest: intelligent agricultural and forestry equipment, Email: qchen@njfu.edu.cn.

***Corresponding author: Jinpeng Wang**, Associate Professor, research interest: agricultural and forestry machinery and automation. School of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing, Jiangsu 210037, China. Tel: +86-15951705992, Email: jpwang@njfu.edu.cn; **Song Mei**, Associate Researcher, research interest: agricultural mechanization engineering. Nanjing Institute for Agricultural Mechanization. Ministry of Agriculture and Rural Affairs, Nanjing 210014, China. Tel: +86-15366092940, Email: meisong@caas.cn.

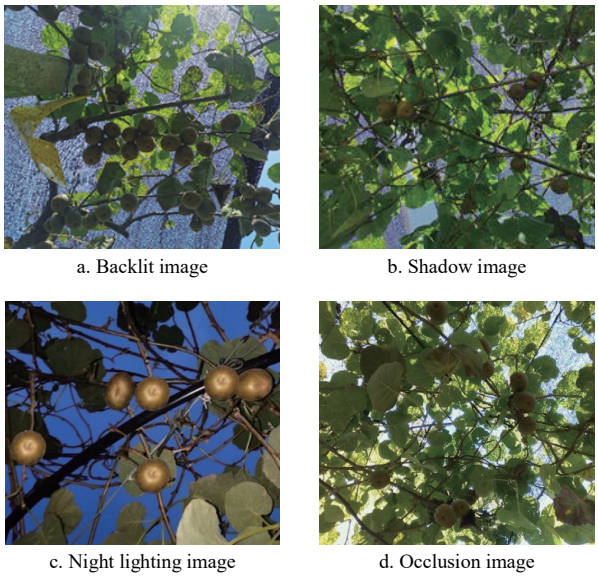
87.89%. Liu et al.[23] proposed a YOLO-Tomato model based on YOLOv3 for tomato image detection, which improved the detection speed and ensured better recognition accuracy. Wang et al.[24] replaced the backbone network of YOLOv4 with MobileNetV3 for dragon fruit recognition and localization; these changes improved the speed of fruit detection, and there was minimal loss of recognition accuracy.

In order to improve the detection ability of the network for kiwifruit in the natural environment, this study proposed a lightweight convolutional neural network YOLOv4-GS algorithm for kiwifruit detection based on YOLOv4. By replacing the backbone network CSPDarknet-53 with GhostNet, the backbone network of YOLOv4 was used to detect kiwifruit. It improved the detection accuracy of the network for small target kiwifruit. At the same time, the upsampling feature map fusion was performed on the 151st and 154th layers of the network, and the Spatial Pyramid Pooling (SPP) network was removed to reduce the amount of network calculation and improve the detection speed of kiwi-fruit. Finally, the network is deployed to the edge device. Experiments show that the YOLOv4-GS algorithm has a good effect on improving the detection accuracy and speed of kiwifruit. It lays a theoretical foundation for automatic kiwifruit picking in complex environments.

2 Materials and methods

2.1 Image acquisition

Based on the growth characteristics and cultivation patterns of kiwifruits, images of kiwifruits were obtained from a vertical upward view, as shown in Figure 1. The images were collected from September to October 2021 at the Lile agricultural plantation in Nanjing. The kiwifruit variety was Hongyang. To ensure the diversity of kiwifruit images, the images were collected at different times of the day and under different weather conditions. Finally, 1500 kiwifruit images were obtained with a resolution of 640×480 pixels; these images were saved in JPEG format. Because of the growth habit of kiwifruit, several kiwifruit are often crowded together. In addition, the leaf size of kiwifruit is equivalent to that of kiwifruit, so there are more than half of the images with different degrees of occlusion and overlap.



a. Backlit image b. Shadow image

c. Night lighting image d. Occlusion image

Figure 1 Kiwifruit image examples in natural environments

2.2 Image data expansion

After randomly flipping, randomly rotating, adjusting image brightness and contrast, and motion blurring the original images using OpenCV, 2766 kiwifruit images were obtained after manually eliminating the invalid images. The abovementioned steps are discussed below:

1) Flip: Random flip (horizontal or vertical);

2) Rotate: Random rotation between −180° and +180°;

3) Brightness: Random brightness adjustment from −20% to +20%;

4) Contrast: Reduce contrast by 30%;

5) Noise: The salt and pepper noise is added to 5% of the image pixel;

6) Motion blurring: Motion blur the original image.

2.3 Data set annotation and partitioning

In this study, the LabelImg tool was used to label the kiwifruit images. Kiwifruits in each image were labeled and saved in the XML file format. The kiwifruit images were divided into a training set, a test set, and a validation set in the 7:2:1 ratio. The exact number of datasets is listed in Table 1. In the process of image annotation, considering that the color of kiwifruit leaves is similar to that of kiwifruit, it is more likely to cause false detection when the occlusion is serious. The kiwifruits with serious occlusion are reserved for visual recognition and picking when the robot moves to the next angle. Therefore, the kiwifruits with serious occlusion are not marked in the process of image annotation.

Table 1 Number of images in the datasets

Dataset	Training set	Test set	Validation set	Total
Sunny image	771	220	111	1102
Cloudy image	743	212	107	1062
Night fill light image	422	121	59	602

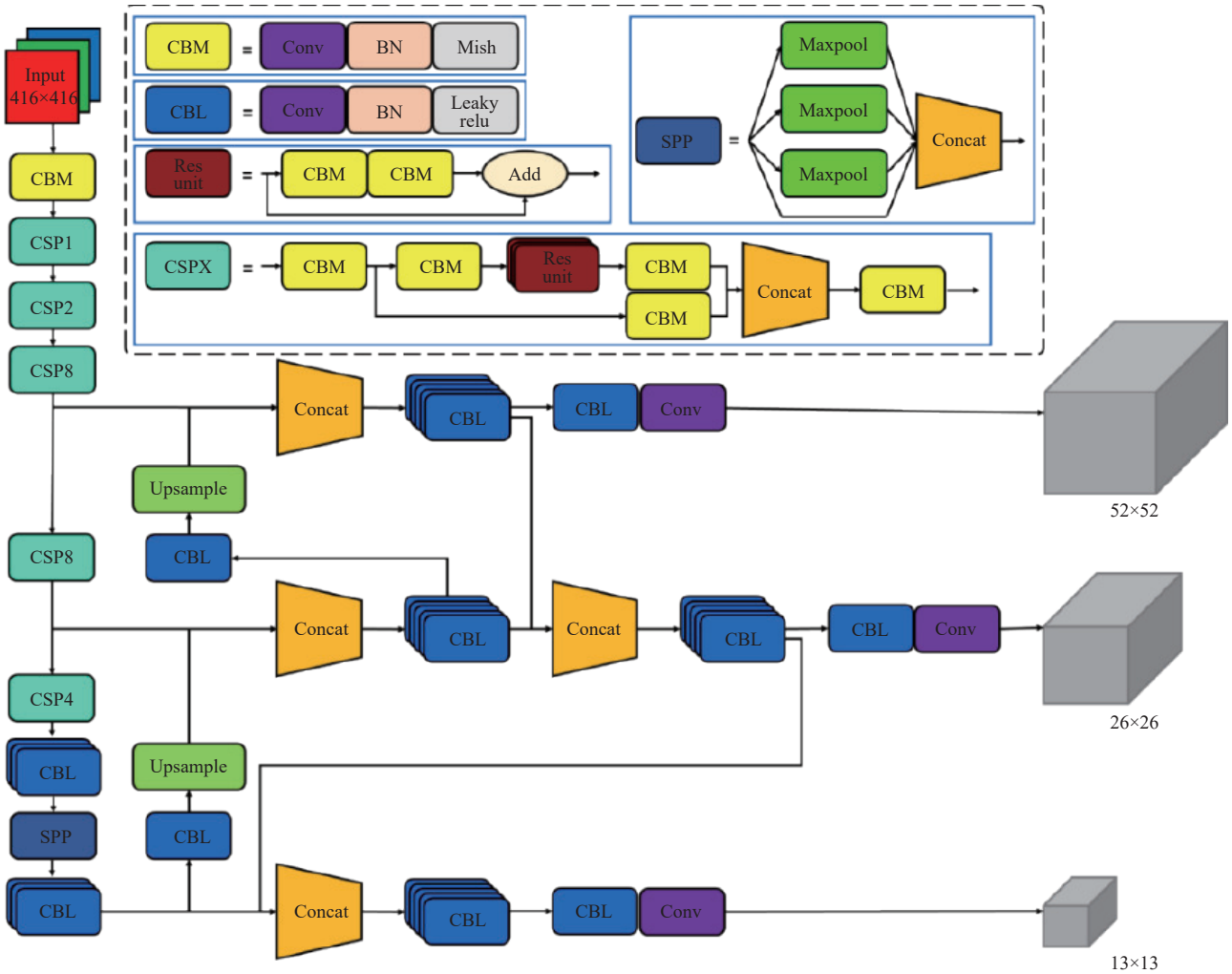
2.4 YOLOv4

YOLOv4 integrates multiple optimization strategies to achieve a perfect balance of detection speed and accuracy[25]. The YOLOv4 model mainly consists of the backbone network CSPDarknet-53, the SPP module, the PANet (Path Aggregation Network) feature map fusion module, and the YOLO head classifier. The complete structure is shown in Figure 2.

As shown in Figure 2, an input image of size 416×416 was taken as an example. CSPDarknet-53 was the feature extraction network used to acquire three preliminary effective feature layers. Compared to Darknet-53, CSPDarknet-53 achieved a richer combination of gradients, and it reduced the computational effort; this enhanced the learning capability of the backbone network and avoided the gradient disappearance problem caused by deeper networks. In addition, the SPP module used four maximum pooling methods to integrate the feature maps of different sizes. The three scales of feature maps obtained after downsampling by the CSPDarknet-53 network were based on the top-down upsampling method used by the feature pyramid network (FPN) for feature fusion; FPN was combined with the PANet network structure to form a down-up feature pyramid again. The final predictive output obtained three feature maps 13×13, 26×26, and 52×52 with different receptive fields to detect large, medium, and small targets, respectively. The results obtained from this research can be applied for the real-time detection of kiwifruits; therefore, the model needed to be lightened and improved to reduce the computational effort as much as possible and increase the detection speed of the model.

2.5 YOLOv4-GS

To improve the detection speed of the model, the backbone network of YOLOv4 was adapted to achieve feature extraction



Note: CBM block includes Conv. layers, BN, and Mish activation function. CBL block includes Conv. layers, BN, and Leaky relu activation function. Res. Unit block is mainly composed of CBM blocks. CSPX (*X* referred to 1, 2, 4, and 8) consists of CBM blocks and Res. Unit blocks. SPP is mainly composed of Maxpools.

Figure 2 Structure of YOLOv4, which consists of backbone network (CSPDarknet53), neck (FPN, PANet, and SSP), and heads (YOLO Heads)

using GhostNet. GhostNet, proposed by Huawei Noah's Ark Lab, was designed for mobile devices and could be carried on embedded devices for real-time detection, which effectively avoided insufficient memory problems and the high latency caused by overly complex models[26]. GhostNet was based on the Ghost bottleneck, which was built on the Ghost Module. The Ghost Module was started because of the problem of feature map redundancy, and it generated multiple feature maps with only a small amount of computation. When compared with the original convolution operation, the Ghost Module could reduce the amount of calculation by half. GhostNet followed the advantages of the basic architecture of MobileNetV3[27]. Then, the Ghost bottleneck was used to replace the bottleneck in MobileNetV3. In the experiment of the ImageNet classification task, the accuracy of GhostNet was higher than that of the MobileNet series[28], ShuffleNet series[29], and FBNet[30]. In terms of hardware inference speed, only GhostNet required less running time to achieve the same accuracy as MobileNetV3.

To reduce the computation of the prediction layer of YOLOv4 and improve the detection accuracy of GhostNet for small targets, YOLOv4-GS retained PANet and FPN and removed the redundant computation of the SPP network. Taking the 416×416 size image input as an example, double upsampling was performed for the output of network layer 151 to fuse the feature map with layer 76. The fusion was followed by a 1×1 convolution to enhance the feature map dimension. Double upsampling was performed on layer 154 to fuse the feature map with layer 36, and a 52×52 scale feature map was obtained after three convolutions for detecting small targets. The feature maps of layers 158 and 154 were fused after double upsampling. A 26×26 scale feature map was obtained after three convolutions for detecting medium-sized targets; the feature map of layer 165 was fused with the feature map of layer 151 after double upsampling, and a 13×13 scale feature map was obtained after three convolutions for detecting large targets. The network structure is shown in Figure 3.

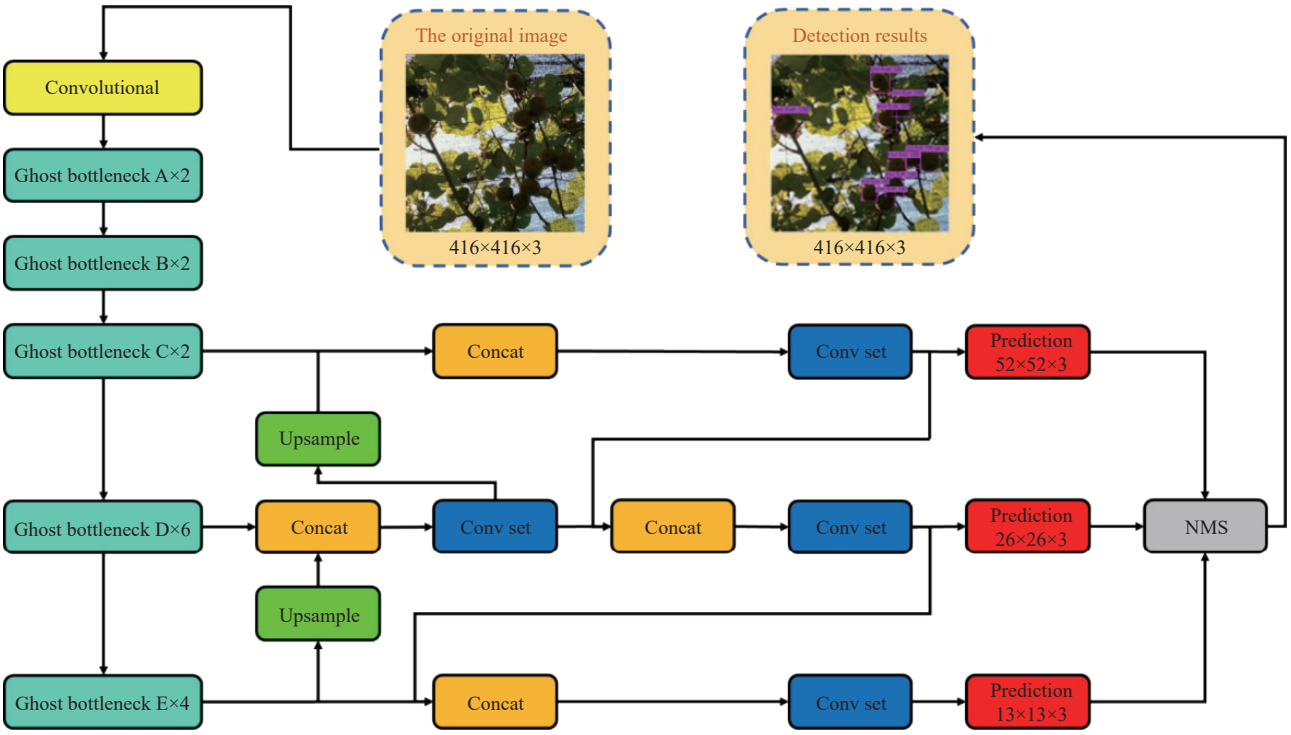
3 Experiment

3.1 Experiment platform

The environment for model training and testing used in this paper was Ubuntu 18.04 operating system, Intel core i5 10400F CPU, 16 GB of memory, NVIDIA Ge Force GTX 1660 Super GPU, CUDA10.2, and CUDNN7.6. In addition, we used Darknet as the deep learning framework for the experiments.

3.2 Experiment parameters

In this study, cluster analysis was performed on the kiwifruit dataset, and the nine anchor frame sizes obtained were (22, 27), (24, 35), (30, 38), (32, 46), (36, 47), (39, 45), (35, 53), (40, 51), and (48, 63). The number of samples per batch was 32, and the



Note: Convolutional is a 3×3 convolution. The Conv Set is a convolution module composed of 1×1, 3×3, and 1×1. The NMS is non-maximum suppression.

Figure 3 Structure of YOLOv4-GS

momentum was 0.95. The decay was 0.0005, and the initial learning rate was 0.001; the number of iterations was 50 000. The learning rate decreased to 0.0001 after 35 000 iterations and to 0.000 01 after 40 000 iterations. The weights file was updated after every 1000 iterations.

3.3 Model evaluation

To analyze the performance of the target detection algorithm, the F1-score, Average Precision (AP), Intersection over Union (IoU), average detection time, and model weight were used as evaluation metrics to assess the comprehensive performance of the kiwifruit recognition model. The F1-score and AP are defined as follows:

$$\mathrm{F1} = 2\frac{P \cdot R}{P + R} \tag{1}$$

$$\mathrm{AP} = \int_0^1 P(R)\mathrm{d}R \tag{2}$$

where, P denotes the accuracy rate, R denotes the recall rate, F1 denotes the equally weighted summed average of the accuracy and recall rates, and AP denotes the average accuracy rate.

IoU represents the accuracy of the target spatial feature prediction in target detection, with good predictions having high IoU values. IoU is defined as follows:

$$\mathrm{IoU} = \frac{S(A \cap B)}{S(A \cup B)} \tag{3}$$

where, A denotes the prediction area of the algorithm, B is the real area of the target, and S is the area of the region.

4 Results and analysis

4.1 Analysis of experiment results

Figure 4 shows the curve of the loss value with the number of iterations before and after the improvement; 50 000 training iterations were performed with the same parameters. The loss value of YOLOv4-GS decreased quickly, and the final loss value was small. According to the model evaluation metrics, the F1-score, AP, and IoU of YOLOv4-GS were 98.00%, 99.22%, and 88.92%, respectively. The average time taken to detect a 416×416 kiwifruit image was 11.95 ms. Figure 5 shows the visualization of kiwifruit recognition by the improved before-and-after models. Both the improved before-and-after models were effective in recognizing kiwifruits under strong backlit and shaded environments. The improved model not only has a better recognition effect on the occluded kiwi fruit but also reduces false recognitions. Overall, as shown in Figure 6, the improved model showed significant advantages in terms of improved recognition, missed detection (see Figures 6a and 6b), false detection (see Figures 6c and 6d), and small target recognition (see Figures 6e and 6f).

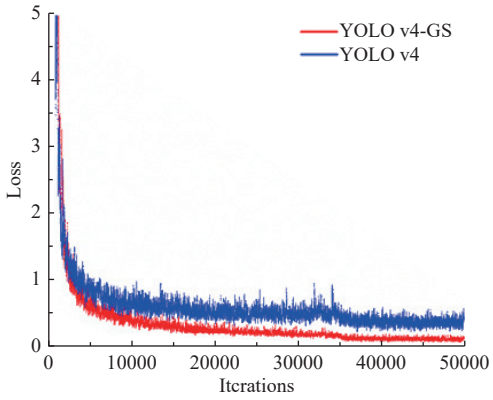
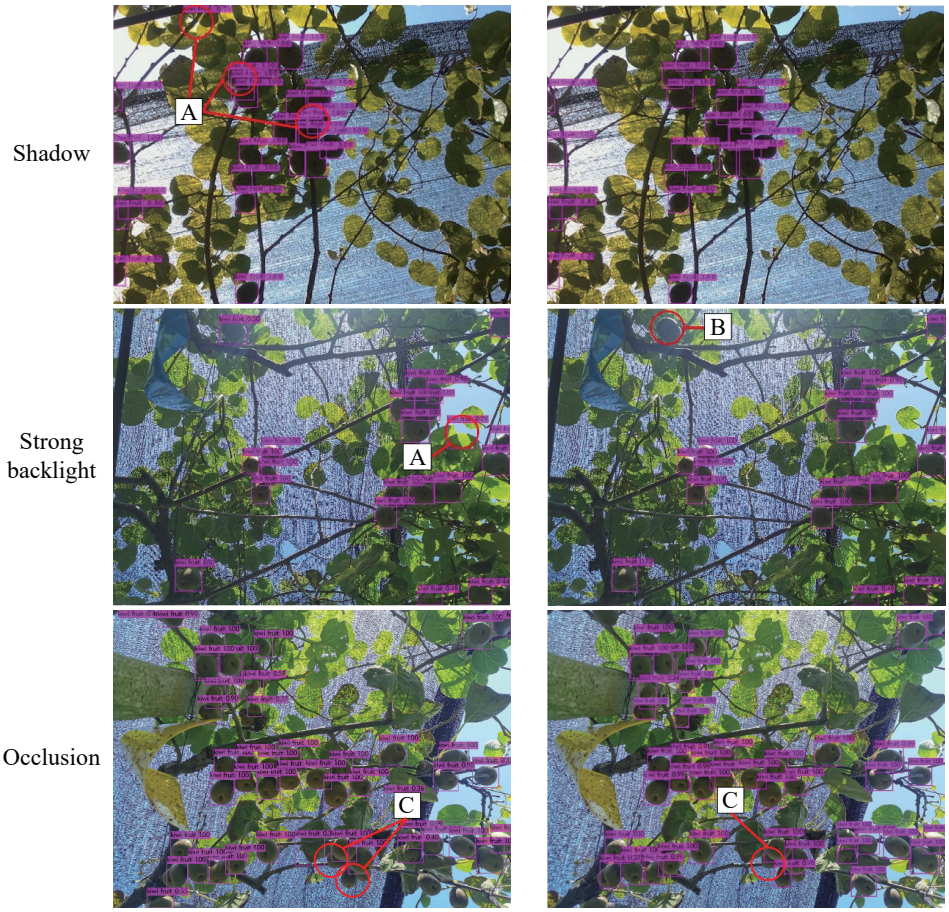


Figure 4 Curve of loss values with the number of iterations

4.2 Analysis of different backbone networks

To objectively analyze the performance of the YOLOv4-GS model, training and analyses were performed for different backbone networks. Darknet-19, Darknet-53, CSPDarknet-53, and GhostNet were selected for the comparative analysis. The performance of the GhostNet backbone network was objectively evaluated by ensuring that the parameters were consistent during training, and the experiment results are listed in Table 2. Darknet-53 had a more



a. YOLO v4 b. YOLO v4-GS

Note: The label "A" with the red circle indicates false detection. The label "B" with the red circle indicates missed detection. The label "C" with the red circle indicates that the blocked kiwifruit is not detected.

Figure 5 Examples of kiwifruit images detected using YOLOv4 and YOLOv4-GS under different environment situations

Table 2 Detection results of different backbone networks

Parameters	Darknet-19	Darknet-53	CSPDarknet-53	GhostNet
Precision	0.96	0.96	0.95	0.97
Recall	0.94	0.99	1.00	0.98
F1-score/%	95.00	97.00	98.00	98.00
AP/%	96.45	98.39	99.29	99.22
IoU/%	84.52	86.09	87.94	88.92
Average Time/ms	32.95	42.56	43.39	11.95
Model weight/MB	202.0	246.3	256.0	28.8

complex network structure than Darknet-19, which resulted in a 1.94% increase in AP and a 1.57% increase in IoU. However, the average detection time of a single image was increased by 9.61 ms. CSPDarknet-53 had higher AP and IoU than Darknet-53, but the detection speed decreased, and the model weight was 256 MB, which made it difficult to meet the requirements of high detection speed on embedded devices. GhostNet had made lightweight improvements on the backbone network. The AP of GhostNet was 0.07% lower than that of CSPDarknet-53, and the detection speed was greatly improved. The average detection time of a single image was 31.44 ms lower than that of CSPDarknet-53. In addition, the weight of YOLOv4-GS was 28.8 MB, which was 227.2 MB less than that of CSPDarknet-53; this weight reduction greatly decreased the operating cost of the embedded devices. By comparing and analyzing different backbone networks, it was found that GhostNet had the characteristics of high detection accuracy, fast detection speed, and low model memory consumption, which could be applied for the real-time detection of kiwifruits in their natural environments and were obviously advantageous in embedded devices.

4.3 Analysis of different network models

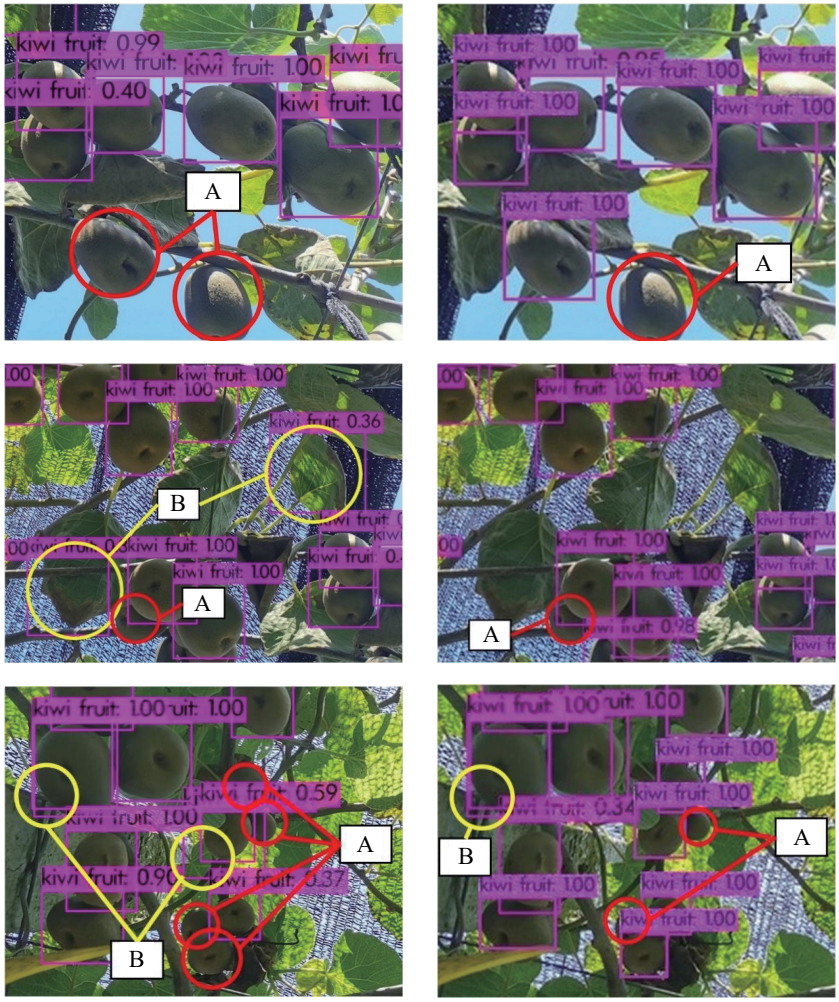
To objectively analyze the performance of YOLOv4-GS, it was compared and analyzed with different network models. This ensured that the parameters were consistent during the training; the curve of the loss with the iterations is shown in Figure 7 and the experiment results are listed in Table 3. The AP of YOLOv4-GS was 8.39% higher than that of Faster R-CNN and 8.36% higher than that of SSD-300. The detection speed of YOLOv4-GS was 11.3 times higher than that of Faster R-CNN and 2.6 times higher than that of SSD-300. The detection results of the three models are shown in Figure 8. Faster R-CNN and SSD-300 had numerous missed detections in identifying kiwifruits in their natural environments, whereas YOLOv4-GS had very few missed detections.

Table 3 Detection results of different network models

Network models	Precision	Recall	F1-score/%	AP/%	Average time/ms
Faster R-CNN	0.65	0.99	78.00	90.83	135.167
SSD-300	0.61	0.97	75.00	90.86	30.74
YOLOv4-GS	0.97	0.98	98.00	99.22	11.95

Note: AP: Average Precision.

In this study, YOLOv4-GS was compared with the models proposed in the studies using YOLO-Tomato[23], YOLOv3-dense[31], R-FCN[32], and Im-AlexNet[33]. The comparison results are listed in Table 4. The F1-score of YOLOv4-GS was 4.09%, 16.00%, and



a. YOLO v4 b. YOLO v4-GS

Note: The label "A" with a red circle indicates missed detection. The label "B" with a yellow circle indicates false detection.

Figure 6 Comparison of missed and false detections between YOLOv4 and YOLOv4-GS

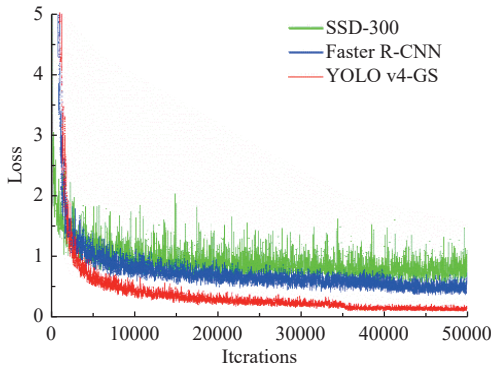


Figure 7 Curve of loss values with the number of iterations

8.00% higher than the F1-scores of YOLO-Tomato, YOLOv3-dense, and R-FCN, respectively. The AP values of the YOLOv4-GS were 2.82%, 4.12%, and 3.22% higher than the AP values of YOLO-Tomato, R-FCN, and Im-AlexNet, respectively. In terms of the detection speed, the YOLOv4-GS showed excellent performance. The average detection time of each YOLOv4-GS image was 42.05 ms faster than YOLO-Tomato, 27.6 times faster than that of the YOLOv3-dense, 15.6 times faster than that of the R-FCN, and 89.5 times faster than that of the Im-AlexNet. A comparison of the different network models shows that YOLOv4-GS was superior to other networks in terms of both detection accuracy and detection speed.

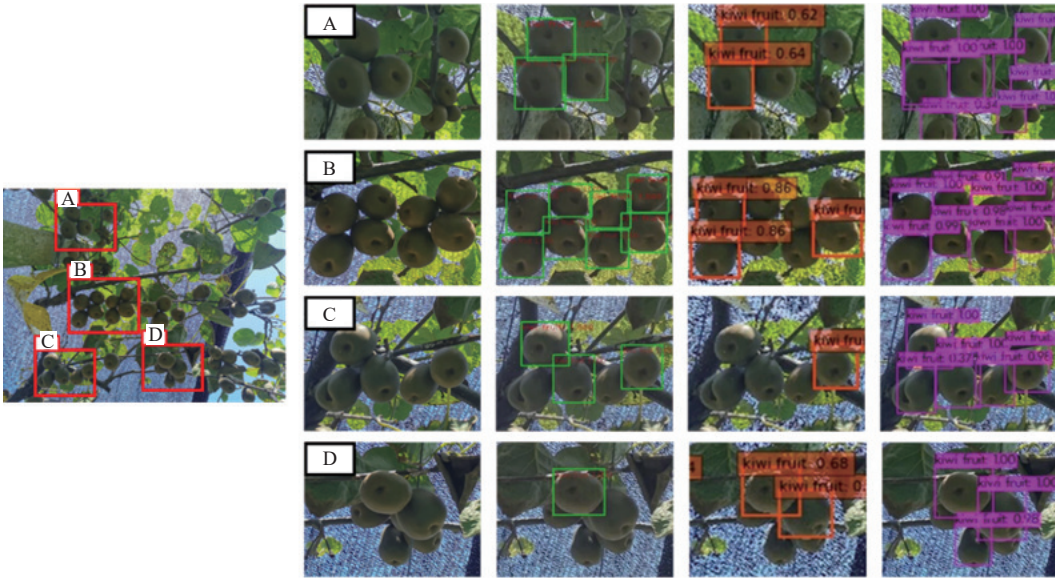
Table 4 Comparison of different network models

Network models	F1-score/%	AP/%	Average time/ms	Detection object
YOLO-tomato	93.91	96.40	54.00	Tomato
YOLOv3-dense	82.00	--	330.00	Apple
R-FCN	90.00	95.10	187.00	Apple
Im-AlexNet	--	96.00	1070.00	Kiwifruit
YOLOv4-GS	98.00	99.22	11.95	Kiwifruit

4.4 Kiwifruit picking experiment and analysis

In order to further analyze the real-time recognition effect of the model proposed in this paper in embedded equipment, indoor picking experiments and orchard picking experiments were carried out in the laboratory and the plantation of Nanjing Lile Agricultural Company in October 2021, as shown in Figure 9. The experimental equipment includes a ZED binocular stereo camera, Jetson Xavier NX system, and S6H4D_Plus six-axis manipulator. The experimental object was Hongyang kiwifruit. The flow chart of the picking kiwifruits test is shown in Figure 10.

The kiwifruits were arranged in separate, adjacent, and sheltered conditions for indoor experiments. There were 21 kiwifruits divided into nine groups. The experimental results showed that the average speed of video processing of YOLOv4-GS deployed to the embedded system reaches 28.4 fps, and the average time of picking a single kiwifruit was 11.72 s. The model proposed in this paper has a good effect on the accuracy and speed of kiwifruit recognition.



Note: Label "A" represents the detection of small target kiwifruits. Labels "B" and "C" represent the detection of dense clusters of kiwifruits. Label "D" represents the detection of obscured kiwifruits.

Figure 8 Detection results of the three models

Note: 1. Indoor kiwifruits scaffolding;
2. S6H4D_Plus six-axis manipulator;
3. Identification interface;
4. ZED binocular stereo camera;
5. Collection box; 6. Jetson Xavier NX system
a. Indoor picking experiment

b. Orchard picking experiment

Figure 9 Experiments of picking kiwifruits

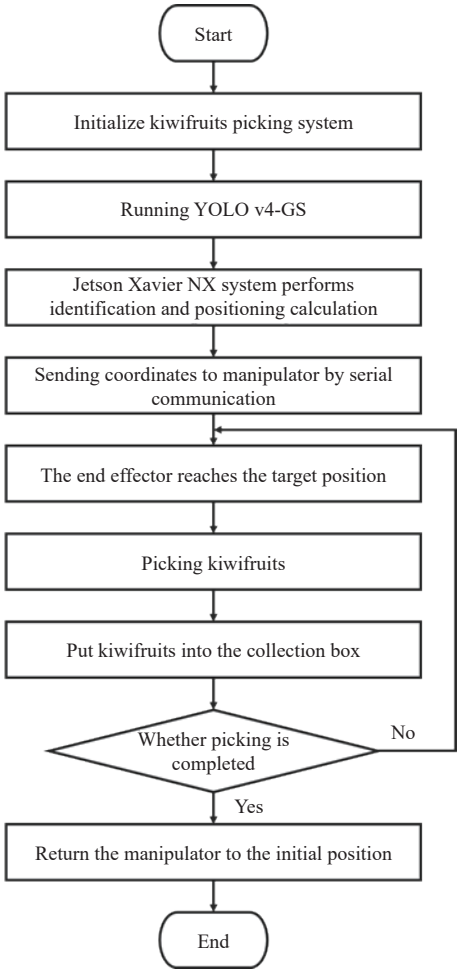


Figure 10 Flow chart of picking kiwifruits test

A total of 50 kiwifruits were picked in the orchard picking experiment. The recognition accuracy, positioning accuracy, picking success rate, and overall evaluation success rate of the kiwifruit picking system were about 90.00%, 95.60%, 90.69%, and 78.00%, respectively. The specific error factors in the picking process are listed in Table 5. The time-consuming process of picking the experiment in the natural environment is listed in Table 6. The average time-consuming of recognition and positioning was 6.09 s, accounting for about 29.03% of the total time. In conclusion, the recognition and positioning method based on YOLOv4-GS satisfied the efficiency and accuracy requirements of the kiwifruit

picking system. In practical application, kiwifruit with a small occlusion area can be effectively recognized and successfully picked, but kiwifruit with serious occlusion is not recognized for the first time. With the movement of the picking robot, the angle of view changes, and the occlusion area becomes smaller. After the kiwifruit is successfully recognized, the picking is completed.

Table 5 Analysis of orchard picking experiment results

Total errors	Error type	Errors/Total	Cause
11	Recognition error	5/50	Severe shielding, strong backlight, or dark environment.
	Positioning deviation	2/45	Fruit shaking, strong light, or uneven light.
	Picking error	4/43	The end effector width is insufficient. The position of the fruit exceeds the working range of the manipulator.

Table 6 Time of different actions during the orchard picking experiment

Action	Average time/s	Proportion/%
Identification and location	6.09	29.03
Manipulator movement	6.36	30.31
Picking	2.01	9.58
Manipulator reset	6.52	31.08
Total	20.98	100

5 Conclusions

This study explored the most efficient method for kiwifruit detection in natural environments. The YOLOv4-GS proposed in this study used a GhostNet feature extraction network and upsampled feature map fusion on the network layers 151 and 154. The SPP network was removed to improve the model detection speed while the detection accuracy was maintained. The experiment results proved that the F1-score, AP, and IoU of YOLOv4-GS were 98.00%, 99.22%, and 88.92%, respectively. The average detection time of YOLOv4-GS was 31.44 ms less than that of YOLOv4, and the model weight was reduced by 227.2 MB, which led to a significant improvement in the detection speed. In addition, YOLOv4-GS had a significantly improved detection accuracy and speed than Faster R-CNN and SSD-300. In the indoor picking experiment and the orchard picking experiment, the average speed of the YOLOv4-GS processing video was 28.4 fps. The recognition accuracy was above 90%. The average time spent for recognition and positioning was 6.09 s, accounting for about 29.03% of the total picking time.

Overall, YOLOv4-GS performs well with good detection capability and can quickly and accurately identify kiwifruit in complex environments. However, under the condition of too strong or too dark light, the kiwifruit will miss recognition. In the future, when picking outdoors, the influence of natural light on recognition can be reduced by adding a filter to the camera. In addition, in the future, it is necessary to re-shoot the data set for model training and keep the distance between the camera and the camotea tree in the data set consistent with the actual harvest distance. Some blurred images should be added to the data set to enrich its diversity so that the trained model can adapt to the recognition of kiwifruits under different lighting conditions.

Acknowledgments

This research was funded by the Jiangsu Province Agricultural Science and Technology Independent Innovation Project (CX(22)3099), the Emergency Science and Technology Project of National Forestry and Grassland Administration (202202-3), the Key R&D Program of Jiangsu Modern Agricultural Machinery Equipment and Technology Promotion Project (Grant NJ2021-18), the Key R&D plan of Jiangsu Province (Grant BE2021016-2), and the 2021 Self-made Experimental Teaching Instrument Project of Nanjing Forestry University (Grant nlzzyq202406).

[References]

[1] Xiao X, Li M. Fusion of data-driven model and mechanistic model for kiwifruit flesh firmness prediction. Computers and Electronics in Agriculture, 2022; 193: 106651.

[2] Yang C, Lee W S, Gader P. Hyperspectral band selection for detecting different blueberry fruit maturity stages. Computers and Electronics in Agriculture, 2014; 109: 23–31.

[3] Song Z Z, Zhou Z X, Wang W Q, Gao F F, Fu L S, Li R, et al. Canopy segmentation and wire reconstruction for kiwifruit robotic harvesting. Computers and. Electronics in Agriculture, 2021; 181: 105933.

[4] Mu L T, Liu H Z, Cui Y J, Fu L S, Gejima Y. Mechanized technologies for scaffolding cultivation in the kiwifruit industry: A review. Information Processing in Agriculture, 2018; 5(4): 401–410.

[5] Zhang Z, Igathinathane C, Li J, Cen H, Lu Y, Flores P. Technology progress in mechanical harvest of fresh market apples. Computers and Electronics in Agriculture, 2020; 175: 105606.

[6] Jia W K, Zhang Z H, Shao W J, Hou S J, Ji Z, Liu G L, et al. FoveaMask: A fast and accurate deep learning model for green fruit instance segmentation. Computers and Electronics in Agriculture, 2021; 191: 106488.

[7] Cui Y J, Su S, Wang X X, Tian Y F, Li P P, et al. Recognition and feature extraction of kiwifruit in natural environment based on machine vision. Transactions of CSAM, 2013; 44(5): 247–252. (in Chinese)

[8] Tian K, Li J H, Zeng J F, Evans A, Zhang L N. Segmentation of tomato leaf images based on adaptive clustering number of *K*-means algorithm. Computers and Electronics in Agriculture, 2019; 165: 104962.

[9] Maldonado W, Barbosa J C. Automatic green fruit counting in orange trees using digital images. Computers and Electronics in Agriculture, 2016; 127: 572–581.

[10] Wiatowski T, Bölcskei H. A mathematical theory of deep convolutional neural networks for feature extraction. IEEE Transactions on Information Theory, 2018; 64(3): 1845–1866.

[11] Majeed Y, Karkee M, Zhang Q. Estimating the trajectories of vine cordons in full foliage canopies for automated green shoot thinning in vineyards. Computers and Electronics in Agriculture, 2020; 176: 105671.

[12] Zhou J, Fu X Q, Zhou S Q, Zhou J F, Ye H, Nguyen H T. Automated segmentation of soybean plants from 3D point cloud using machine learning. Computers and Electronics in Agriculture, 2019; 162: 143–153.

[13] Xu C Y, Liu Y, Ding F L, Zhuang Z L. Recognition and grasping of disorderly stacked wood planks using a local image patch and point pair feature method. Sensors, 2020; 20(21): 6235.

[14] Jin X J, Sun Y X, Yu J L, Chen Y. Weed recognition in vegetable at seedling stage based on deep learning and image processing. Journal of Jilin University (Engineering and Technology Edition), 2023; 53(8): 2421–2419. (in Chinese)

[15] Ni C, Li Z Y, Zhang X, Zhao L, Zhu T T, Jiang X S. Film sorting algorithm in seed cotton based on near-infrared hyperspectral image and deep learning. Transactions of the CSAM, 2019; 50(12): 170–179. (in Chinese)

[16] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014; pp.580–587. doi: 10.1109/CVPR.2014.81.

[17] Ren S Q, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015; 39(6): 1137–1149.

[18] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y, Berg A C. SSD: Single shot multibox detector. In: Proceedings of the European Conference on Computer Vision – ECCV 2016, 2016; pp.21–37. doi: 10.1007/978-3-319-46448-0_2.

[19] Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. In: 2016 IEEE Conference on

Computer Vision and Pattern Recognition (CVPR), Las Vegas: IEEE, 2016; pp.779–788. doi: 10.1109/CVPR.2016.91.

[20] Xiong J T, Liu Z, Tang L Y, Lin R, Bu R B, Peng H X. Visual detection technology of green citrus under natural environment. Transactions of the CSAM, 2018; 49(4): 45–52. (in Chinese)

[21] Song Z Z, Fu L S, Wu J Z, Liu Z H, Li R, Cui Y J. Kiwifruit detection in field images using Faster R-CNN with VGG16. IFAC-PapersOnLine, 2019; 52(30): 76–81.

[22] Li S J, Hu D Y, Gao S M, Lin J H, An X S, Zhu M. Real-time classification and detection of citrus based on improved single short multibox detecter. Transactions of the CSAE, 2019; 35(24): 307–313. (in Chinese)

[23] Liu G X, Nouaze J C, Touko Mbouembe P L, Kim J H. YOLO-Tomato: A robust algorithm for tomato detection based on YOLOv3. Sensors, 2020; 20(7): 2145.

[24] Wang J P, Gao K, Jiang H Z, Zhou H P. Method for detecting dragon fruit based on improved lightweight convolutional neural network. Transactions of CSAE, 2020; 36(20): 218–225. (in Chinese)

[25] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal speed and accuracy of object detection. arXiv 2020. arXiv: 2004.10934.

[26] Han K, Wang Y H, Tian Q, Guo J Y, Xu C J, Xu C. GhostNet: More features from cheap operations. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020; pp.1577–1586. doi: 10.1109/CVPR42600.2020.00165.

[27] Howard A, Sandler M, Chen B, Wang W J, Chen L-C, Tan M X, et al. Searching for MobileNetV3. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul: IEEE, 2019; 1314–1324. doi: 10.1109/ICCV.2019.00140.

[28] Howard A G, Zhu M L, Chen B, Kalenichenko D, Wang W J, Weyand T, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv 2017, arXiv: 1704.04861.

[29] Zhang X Y, Zhou X Y, Lin M X, Sun J. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City: IEEE, 2018; pp.6848–6856. doi: 10.1109/CVPR.2018.00716.

[30] Wu B C, Dai X L, Zhang P Z, Wang Y H, Sun F, Wu Y M, et al. FBNet: Hardware-aware efficient ConvNet design via differentiable neural architecture search. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019; pp.10726–10734. doi: 10.1109/CVPR.2019.01099.

[31] Tian Y N, Yang G D, Wang Z, Wang H, Li E, Liang Z Z. Apple detection during different growth stages in orchards using the improved YOLO-V3 model. Computer and Electronics in Agriculture, 2019; 157: 417–426.

[32] Wang D D, He D J. Recognition of apple targets before fruits thinning by robot based on R-FCN deep convolution neural network. Transactions of the CSAE, 2019; 35(3): 156–163. (in Chinese)

[33] Mu L T, Gao Z B, Cui Y J, Li K, Liu H Z, Fu L S. Kiwifruit detection of far-view and occluded fruit based on improved AlexNet. Transactions of the CSAM, 2019; 50(10): 24–34. (in Chinese)