DEEP LEARNING-BASED TARGET RECOGNITION AND LOCALIZATION IN CITRUS PICKING ROBOT
/
基于深度学习的柑橘采摘机器人目标识别与定位方法研究

Yu YANG[*1)], Linhui WANG[1)], Shuwei DENG[1)], Zhengqi ZHOU[1)], Zhizhuang LIU[1)]
[1)] Hunan Institute of Science and Engineering, Yongzhou, Hunan 425199, China
E-mail: u7e840@yeah.net
DOI: https://doi.org/10.35633/inmateh-79-44

Keywords: *deep learning, citrus, target recognition, depth camera*

ABSTRACT
Target recognition and localization of picking robots are very important in orchard environments. This paper analyzed the You Only Look Once version 8 normal (YOLOv8n) method in deep learning for citrus picking robots, introduced FasterNet, efficient multi-scale attention (EMA) mechanism, and Wise Intersection over Union (WIoU) loss function, developed an improved YOLOv8n method, and collected a citrus fruit dataset to analyze the recognition and positioning effects of the proposed methods. The selected FasterNet, EMA, and WIoU were all superior to the other types compared, and the improved YOLOv8n method achieved an accuracy of 0.893, a recall rate of 0.894, and a mean average percentage (mAP) of 0.895 for citrus recognition, outperforming other recognition methods. In terms of citrus localization, the average errors on X, Y, and Z axes were 3.56 mm, 3.54 mm, and 4.74 mm, respectively. The findings demonstrate the reliability of the proposed model, and it can be applied to actual citrus picking robots.

摘要
在果园环境下，采摘机器人的目标识别与定位十分重要。本文针对柑橘采摘机器人，分析了深度学习中的YOLOv8n 方法，引入 FasterNet、EMA 注意力机制以及 WIoU 损失函数，设计了一种改进 YOLOv8n，并采集柑橘果实数据集，对所提方法的识别和定位效果进行了分析。结果发现，所选择的 FasterNet、EMA 以及 WIoU 均优于所比较的其他类型，改进 YOLOv8n 在柑橘识别上获得了 0.893 的准确率、0.894 的召回率和 0.895 的 mAP，优于其他识别方法。在柑橘定位上，在 X、Y、Z 上的平均误差分别为 3.56mm、3.54mm 和 4.74mm。结果证明了所提方法的可靠性，可以在实际的柑橘采摘机器人中进行应用。

INTRODUCTION
With the advancement of technologies such as intelligent control and mechanized equipment, the agricultural industry has also undergone tremendous changes. In the fruit industry, due to the limitations of planting patterns and terrain conditions, the degree of mechanization in orchards is still relatively low. Especially in fruit picking, as the seasonal intensity is high and the picking environment is complex, the labor cost is constantly rising. The high labor cost poses a huge challenge to the development of the fruit industry. The research, development, and application of intelligent picking equipment have become an important aspect of the development of the forest and fruit industry. The picking robot is a type of robot used to complete the picking of fruits and vegetables (*Wang et al., 2024*), and it is an important part of the current research on agricultural intelligence. The machine vision system is the perceptual core of a picking robot, and its target recognition accuracy, real-time inference speed, and three-dimensional positioning accuracy can directly affect the picking success rate and operation stability of the robot. In recent years, deep learning provides new technologies for visual perception in complex agricultural environments with its excellent feature extraction and autonomous learning abilities. *Chen et al.* (*2025*) designed a method for target recognition of agricultural picking robots in low-light conditions using template matching, achieving a maximum recognition accuracy of 98.56%, which enables efficient picking in low-light conditions. *Rong et al.* (*2024*) designed a tomato-picking robot system, and field trials showed that the robot had a harvesting success rate of over 50% and an average harvesting cycle of 24 seconds. *Zhao* (*2024*) developed a You Only Look Once version 5 (YOLOv5)-RACF algorithm for the problem of automatic apple picking, with an average detection accuracy of 98.748% and a time of 0.13 seconds to calculate the diameter of one apple. *Zhu et al.* (*2025*) designed an improved You Only Look Once version 7 (YOLOv7) network for the detection of Agaricus bisporus, and the experiment found that the acquisition success rate of this method was 80%, which could achieve automated acquisition.

Matache et al. (2026) designed a greenhouse tomato harvesting system that uses a hybrid vision framework that integrates convolutional neural networks (CNNs) and watershed-based segmentation to detect and position fruits. This system achieved a fruit detection precision of 96.9%. Citrus is a kind of fruit with high yield and demand at present. In the research of citrus-picking robots, *Lin et al.* (*2019*) designed an algorithm based on red-green-blue-depth (RGB-D) image analysis to detect and localize citrus. They conducted experiments on a dataset of 506 RGB-D images and found that the proposed algorithm was robust, with an F1 score of 0.9197. *Yin et al.* (*2023*) proposed an integrated citrus-harvesting robot solution, which adopted a multi-sensor-based fused simultaneous localization and mapping algorithm to achieve positioning and navigation for the orchard robot. Its overall success rate was 87.2%, and the average picking time was 10.9 seconds per fruit. *Zhang et al.* (*2025*) selected YOLOv5 for citrus object detection and achieved an accuracy of 92.3%, a recall of 79.1%, and a mean average percentage (mAP) of 88.5%, which verifies that this network is capable of handling detection tasks in real environments. *Hou et al.* (*2024*) combined a convolutional block attention module (CBAM) with the backbone network of a Mask region-based convolutional neural network (R-CNN). They found that the mean absolute error and relative error of citrus picking point localization was 8.63 mm and 2.76%, respectively, providing a valid solution for the accurate detection and localization of citrus picking points for harvesting robots.

In the current research, the main deficiencies are as follows. Firstly, most of the existing high-precision algorithms rely on deep and heavy network structures, which have a large number of parameters and high computational requirements, resulting in high hardware deployment costs and making it difficult to meet the real-time operation requirements of picking robots. Secondly, the training and performance verification of most models rely on small-scale and single-scene datasets, leading to insufficient robustness. Thirdly, most existing models have problems such as insufficient feature extraction, weak ability to recognize difficult samples, and low boundary regression accuracy, which limit the practical application of citrus picking robots. In response to the above research status and deficiencies, this paper aimed at high precision and strong robustness and took You Only Look Once version 8 normal (YOLOv8n), which has few parameters and is suitable for embedded devices, as the basic model for design. The effectiveness of the proposed method was verified through training with multi-scene datasets and laboratory positioning experiments, providing theoretical support for the design and application of citrus picking robots.

MATERIALS AND METHODS

Citrus identification model based on YOLOv8

In actual orchards, the robot first needs to accurately identify the target fruit, a process usually accomplished using object detection techniques. Conventional object detection techniques-including R-CNN and its improved variant, Faster R-CNN-have a slower detection speed and are not suitable for citrus picking tasks that require rapid object detection. CNNs are a typical method in deep learning and have shown excellent performance in visual information processing (*Shao et al., 2023*). You Only Look Once (YOLO) is a CNN-based object detection method that can classify and locate images through a single neural network and perform object detection better and faster. Currently, the YOLO series of methods have good applications in many fields such as pedestrian detection and equipment fault identification.

You Only Look Once version 8 (YOLOv) is one of the highly flexible and scalable models in the YOLO series and has been utilized in various domains such as medicine and agriculture (*Jraba et al., 2025*). Among them, YOLOv8n has the characteristics of fewer parameters and smaller size, which is more suitable for resource-constrained environments and has higher real-time performance. Therefore, this paper selected YOLOv8n to develop a citrus recognition model. The main current flaws of YOLOv8n are as follows.

(1) Convolutional mapping operations in the backbone network result in a large computational load for the model;

(2) The influence of overlapping citrus fruits and light intensity in the orchard environment results in the loss of target feature information, and more attention should be paid to key target areas;

(3) The original loss function may cause significant deviations in scenarios such as occlusion, density, and light variations.

Therefore, in order to achieve better results in citrus recognition and enhance the picking performance of robots, improvements need to be made based on the YOLOv8n.

Citrus recognition model based on YOLOv8n

The backbone network of YOLOv8n was improved by introducing FasterNet to reduce computational complexity. FasterNet is a model based on partial convolution (PConv). PConv refers to performing the convolution operation only on some channels of the input feature map while keeping the remaining channels unaltered, thereby reducing the computational load without compromising feature expression ability.

The FasterNet module is shown in Figure 1.

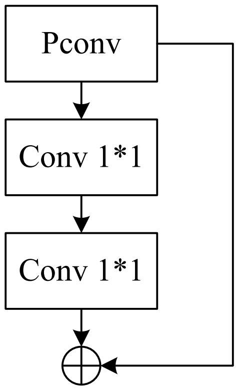

Fig. 1 - FasterNet module

The bottleneck module in the C2f module in the YOLOvn8 backbone network is substituted by the FasterNet module, thereby reducing the computational burden of the model while maintaining high feature expression ability.

By introducing the attention mechanism, the model can focus on information more relevant to citrus recognition among the numerous input information, thereby improving the performance of citrus recognition. Efficient multi-scale attention (EMA) is a mechanism that does not require dimensionality reduction, which significantly reduces computational costs while effectively preserving the integrity of feature channel information. Integrating it into citrus recognition models can enable the models to achieve better performance. The main structure of EMA is shown in Figure 2. EMA is combined with the C2f module to strengthen the model’s capability to extract citrus features.

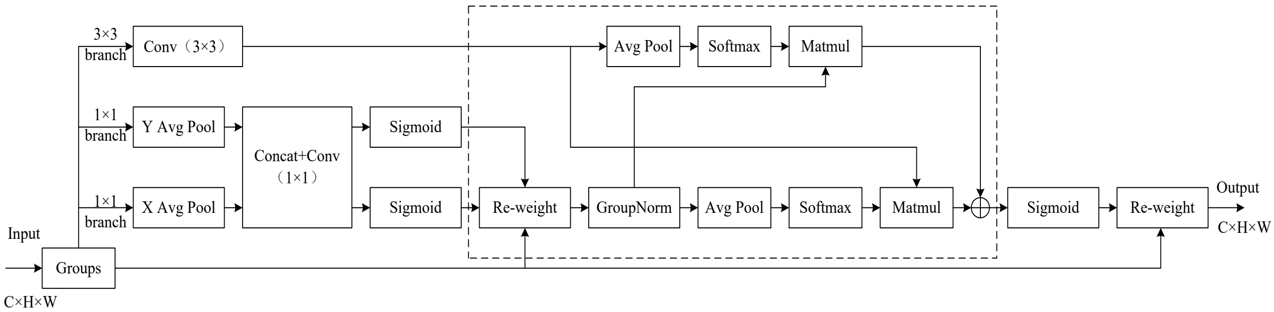

Fig. 2 - EMA mechanism

Finally, Wise Intersection Over Union (WIoU) replaces the Complete Intersection over Union loss (CIoU) function in the original YOLOv8n model.

$WIoU$ is defined as:

$$WIoU=\frac{\sum_{i=1}^{n} w_i IoU(b_i,g_i)}{\sum_{i=1}^{n} w_i} \tag{1}$$

where:

n denotes the number of ground-truth boxes, b_i is the coordinates of the i-th ground-truth box, and g_i is the coordinates of the ground-truth annotation box of the i-th object.

Combining the above three parts, the structure of the improved YOLOv8n is presented in Figure 3.

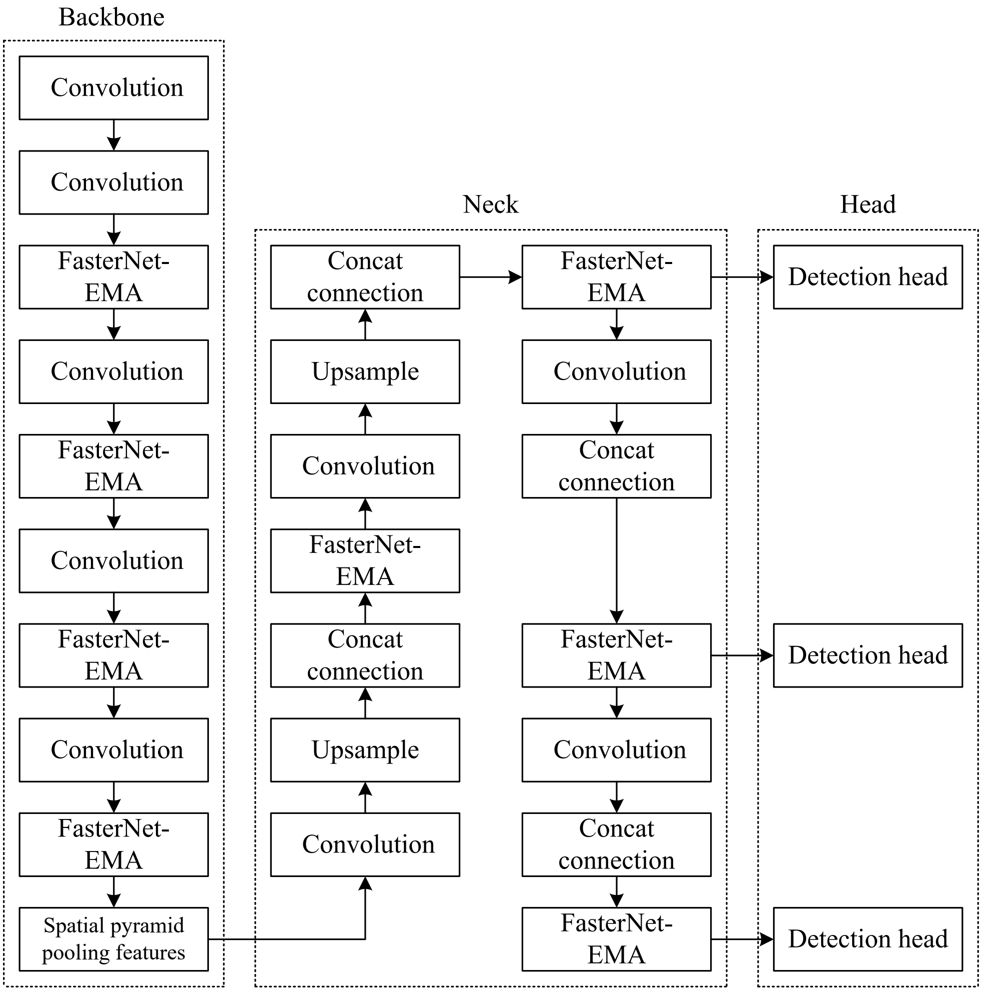

Fig. 3 - The structure of the improved YOLOv8n

Citrus positioning model

During the visual positioning process of the picking robot, a series of coordinate transformations are required to achieve three-dimensional positioning of the recognized fruit. The improved YOLOv8n outputs the rectangular detection box of a citrus target, extracts the pixel coordinates (u, v) of the center point of the box, and then obtains the relationship between the world coordinate system and the pixel coordinate system through coordinate transformation.

First, it is the conversion from the world coordinate system (X_W, Y_W, Z_W) to the camera coordinate system (X_C, Y_C, Z_C):

$$\begin{bmatrix} X_C \\ Y_C \\ Z_C \\ 1 \end{bmatrix} = \begin{bmatrix} R_{3\times3} & t_{3\times1} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_W \\ Y_W \\ Z_W \\ 1 \end{bmatrix} \tag{2}$$

where R refers to a rotation matrix and t is a translation vector.

Then, (X_C, Y_C, Z_C) is converted to the image coordinate system (x, y):

$$\begin{cases} x = \dfrac{f}{Z_C} x_c \\ y = \dfrac{f}{Z_C} y_c \end{cases} \tag{3}$$

Ignoring image distortion, point P in the pixel coordinate system is (u, v), and the correspondence with (x, y) is:

$$\begin{cases} u = \dfrac{x}{dx} + u_0 \\ v = \dfrac{y}{dy} + v_0 \end{cases} \tag{4}$$

where (u_0, v_0) is the projection coordinate of the optical center in the image coordinate system, dx and dy are scaling factors.

When shooting, this article used the Astra Pro depth camera with parameters shown in Table 1.

Table 1

1Astra Pro depth camera parameters

	Parameter
Depth range	0.6-8 m
Relative accuracy	1 m ± 3 mm
Resolution	19,220 × 1,080
Frame rate	30 frames per second
Support operating system	Android/Windows/Linux

For the calibration method, the currently widely used Zhang Dingyou calibration method (*Zhang, 2000*) was used, and Table 2 presents the results of camera calibration.

Table 2

Results of camera calibration

	Calibration result
Internal parameter matrix	$\begin{bmatrix} 784.1254 & 0 & 0 \\ -0.1524 & 784.5524 & 0 \\ 947.5268 & 486.2145 & 1 \end{bmatrix}$
Radial distortion matrix	$[0.1325 \quad 0.4425 \quad -1.0547]$
Tangential distortion matrix	$[0.0063 \quad -0.0026]$

Laboratory experiment

In a laboratory environment, the measurement accuracy of the localization method proposed in this paper for citrus fruits was verified. A depth camera was fixed on the camera support, and a simulated citrus tree model was used to mimic the canopy structure of a real citrus tree. Several citrus fruits were randomly placed on the tree to simulate the picking scenario in an orchard. The laboratory lighting was kept stable to avoid the interference of external light changes. An aurora rangefinder was employed to detect the actual depth values of the fruits, and a spirit level was used to measure the actual coordinates of the fruits. Figure 4 shows the experimental setup.

Fig. 4 - Localization experiment

RESULTS

Dataset establishment

A dataset of citrus fruits was collected in an actual orchard environment to simulate the harvesting scenario of a picking robot. Images were taken in scenarios such as overlapping, dense, and occluded under different environments like front light, backlight, sunny days, and cloudy days. The distance between the camera and the outer edge of the canopy layer was 0.5-1.5 m, depending on the size of the fruit tree, and ultimately 1,186 images were taken. The dataset was divided into training, validation, and testing subsets (80%, 10%, and 10%).

The collected citrus fruit images were labeled using GitHub open-source labeling tool LabelImg (*Du et al., 2024*) to obtain xml format files, and Mosaic data augmentation technique was employed to expand the dataset, yielding a training set of 2,846 images, a validation set of 356 images, and a test set of 356 images.

Experimental setup

The experimental environment and software configurations are presented in Table 3.

Table 3

Experimental environment and software configurations

	Version
Central processing unit	Intel (R) Core i5-13500HX@4.70GHz
Graphics processing unit	RTX4060
Deep learning framework	PyTorch 1.8.1
Python	Python 3.10
Compute Unified Device Architecture (CUDA)	CUDA 11.3

During training, the stochastic gradient descent (SGD) optimizer was used, the epoch was set to 300, the initial learning rate was set to 0.01, the number of weight decay times was 0.005, and the batch size was 32.

Analysis of citrus recognition results

In the improved YOLOv8n model, FasterNet replaced the original backbone network. The recognition results of FastNet and several other models after replacing the YOLOv8n backbone network were compared (Table 4).

Table 4

Comparison of various backbone networks

	Accuracy	**Recall**	**mAP@0.5-0.95**
Original bottleneck	0.863	0.847	0.872
efficientViT *(Mu et al., 2025)*	0.871	0.853	0.875
MobileNetV4 *(Qin et al., 2017)*	0.868	0.836	0.874
HGNetV2 *(Deng et al., 2025)*	0.864	0.851	0.876
FasterNet	0.874	0.857	0.884

According to Table 4, when using FasterNet, the accuracy was 0.874, showing an increase of 1.27% compared to the original bottleneck, the recall rate was 0.857, showing an increase of 1.18%, and the mAP was 0.884, showing an increase of 1.37%. Among the comparisons with other models, FasterNet still performed the best. The results suggest that FasterNet's improvements to YOLOv8n is effective.

The recognition effects of various attention mechanisms were compared.

Table 5

Comparison of various attention mechanisms

	Accuracy	**Recall**	**mAP@0.5-0.95**
None	0.863	0.847	0.872
SENet (*Du et al., 2025*)	0.874	0.843	0.874
Convolutional Block Attention Module (CBAM)	0.872	0.845	0.875
EMA	0.876	0.855	0.881

According to Table 5, the performance of the model was enhanced to varying degrees after introducing the attention mechanism to YOLOv8n. Specifically, it was found that compared with SENet and CBAM, EMA was the best choice. Compared with the original YOLOv8n, accuracy was enhanced by 1.51%, recall rate by 0.94%, and mAP by 1.03%. The simultaneous improvement in accuracy and recall rate was achieved, suggesting the best performance in citrus recognition. The finding demonstrates the reliability of combining EMA with YOLOv8n.

The recognition outcomes of various loss functions were compared.

Table 6

Comparison of various loss functions

	Accuracy	Recall	mAP@0.5-0.95
CIoU	0.863	0.847	0.872
Distance Intersection over Union (DIoU) (*Huo et al., 2023*)	0.867	0.851	0.875
Efficient Intersection over Union (EIoU) (*Kumar & Dhanalakshmi, 2024*)	0.862	0.845	0.873
WIoU	0.871	0.849	0.882

According to Table 6, the recognition effect of the CIoU used in the original YOLOv8n was relatively poor. WIoU achieved a 0.93% enhancement in accuracy, a 0.24% enhancement in recall rate, and a 1.15% enhancement in mAP compared to CIoU, indicating that using WIoU as the loss function in YOLOv8n can achieve better citrus recognition results.

Finally, the improved YOLOv8n model was compared with some of the current popular deep learning-based detection methods.

Table 7

Comparison of various detection methods

	Accuracy	Recall	mAP@0.5-0.95
Faster R-CNN	0.836	0.826	0.846
YOLOv5n	0.851	0.843	0.867
YOLOv8n	0.863	0.847	0.872
Improved YOLOv8n	0.893	0.894	0.895

According to Table 7, Faster R-CNN performed poorly in citrus recognition, with a mAP as low as 0.846, indicating poor usability of the method in citrus recognition. The comparison between YOLOv5n and YOLOv8n showed that the recognition effect of YOLOv8n was better, demonstrating the reliability of the improvement to YOLOv8n. Finally, the improved YOLOv8n model achieved a mAP of 0.895 for citrus recognition, which can satisfy the recognition needs of the robot.

Analysis of citrus localization results

A total of ten sets of coordinate samples were counted, and the actual coordinates and positioning coordinates of citrus fruits were compared.

According to Table 8, the average errors in the X, Y, and Z directions for citrus positioning were 3.56 mm, 3.54 mm, and 4.74 mm respectively, indicating a relatively small overall positioning error. These errors may come from camera calibration errors and system hardware errors. But overall, the average error of the positioning results was within 5 mm, which can satisfy the positioning needs of robots in the actual citrus orchard environment.

Table 8

Analysis of citrus location results

	Actual coordinates/mm	Positioning coordinates/mm	Error value/mm
1	(352.18, 214.37, 1526.42)	(348.57, 217.94, 1522.11)	(3.61, 3.57, 4.31)
2	(487.62, 189.25, 1468.37)	(491.18, 185.73, 1463.89)	(3.56, 3.52, 4.48)
3	(296.73, 256.48, 1583.65)	(300.24, 252.95, 1578.84)	(3.51, 3.53, 4.81)
4	(415.89, 172.64, 1427.83)	(412.37, 176.28, 1423.26)	(3.52, 3.64, 4.57)
5	(371.52, 238.76, 1542.19)	(367.98, 242.33, 1537.92)	(3.54, 3.57, 4.27)
6	(523.47, 165.38, 1396.24)	(519.89, 168.92, 1391.73)	(3.58, 3.54, 4.51)
7	(284.35, 269.52, 1624.78)	(287.93, 265.96, 1620.07)	(3.58, 3.56, 4.71)
8	(438.71, 157.93, 1375.46)	(442.32, 154.46, 1371.18)	(3.61, 3.47, 4.28)
9	(309.64, 247.25, 1563.82)	(306.07, 250.78, 1559.33)	(3.57, 3.53, 4.49)
10	(466.28, 181.37, 1448.59)	(469.83, 177.86, 1444.28)	(3.55, 3.51, 4.31)

2

CONCLUSIONS

In this study, an optimized YOLOv8n model was developed for the recognition and localization of citrus fruits by citrus-picking robots. Through the analysis of a series of performance indicators, it was found that after replacing YOLOv8n backbone network with FasterNet, the recognition accuracy and FPS were improved. This indicates that the design of partial convolutions effectively reduces the memory access cost while maintaining the feature expression ability. In the comparison of attention mechanisms, EMA did not adopt the dimensionality reduction operation, which avoided the loss caused by channel information compression. Meanwhile, its parallel multi-scale structure fully utilized the parallel computing power of the GPU, hardly increasing parameter quantity and the computational burden. The comparison of loss functions demonstrated that WIoU exhibited good accuracy and mAP. This may be because WIoU assigns dynamic weights to samples of different qualities, forcing the model to focus more on the boundary regression of low-quality samples, such as those with severe occlusion and overlap, during training. According to the outcomes of the indoor positioning experiment, the positioning error of the proposed method for citrus fruits was within 5 mm.

Although the research has achieved certain results, there are two limitations.

(1) Dataset scale: The collected data is limited to a single orchard. The generalization ability of the model in more extensive scenarios (different production areas, fruit maturity levels, and tree shapes) needs to be verified.

(2) Lack of actual picking experiments: The positioning experiments were performed in a simulated laboratory environment, using model trees and artificially arranged citrus fruits. The picking situation in a real orchard environment was not involved.

Based on the limitations, future research will expand the dataset by collecting image data of different varieties, at different maturity stages, and under different weather conditions in multiple citrus-producing areas to construct a larger-scale and more representative citrus fruit dataset, build a complete prototype system of citrus picking robots to conduct end-to-end picking experiments in actual orchards, and study the task allocation and path planning methods for multiple picking robots in large-scale citrus orchards to achieve intelligent and clustered harvesting management of orchards.

In summary, the citrus recognition and localization method based on the improved YOLOv8n performs excellently in terms of recognition accuracy, inference speed, and positioning error, providing effective technical support for citrus picking robots. Although there are certain limitations in aspects such as the dataset scale, this study verifies the basic feasibility and advantages of the method.

ACKNOWLEDGEMENT

This work was supported by the Scientific Research Project of Hunan Provincial Department of Education, China (Grant No. 23B0761).

REFERENCES

[1] Chen, H., Zhang, Z., Luo, Q., Chen, R., & Zhao, Y. (2025). A Target Image Recognition Method of Agricultural Harvesting Robot using Template Matching in Low Light Environment. *Pakistan Journal of Agricultural Sciences*, *62*(1), 79-90. https://doi.org/10.21162/PAKJAS/25.66

[2] Deng, L., Miao, Z., Zhao, X., Yang, S., Gao, Y., Zhai, C., & Zhao, C. (2025). HAD-YOLO: An Accurate and Effective Weed Detection Model Based on Improved YOLOV5 Network. *Agronomy*, *15*(1), 57. https://doi.org/10.3390/agronomy15010057

[3] Du, Q., Tang, L., & Li, Y. (2024). SCB-YOLOv5: a lightweight intelligent detection model for athletes' normative movements. *Scientific Reports*, *14*(1), 1-11. https://doi.org/10.1038/s41598-024-59218-w

[4] Du, W., Yu, X., Guo, Z., Wang, H., Pu, Z., & Li, C. (2025). Squeeze-and-excitation attention residual learning of propulsion fault features for diagnosing autonomous underwater vehicles. *Journal of Field Robotics*, 42(1), 169-179. https://doi.org/10.1002/rob.22405

[5] Hou, C., Xu, J., & Fang, M. (2024). Detection and localization of citrus picking points based on binocular vision. *Precision Agriculture*, *25*(5), 2321-2355. https://doi.org/10.1007/s11119-024-10169-2

[6] Huo, C., Ma, J., & Tong, Z. (2023). Research on Pedestrian Crossing Intention Prediction Based on Deep Learning. *Proceedings of the 2023 7th International Conference on Electronic Information Technology and Computer Engineering*, 282-287. https://doi.org/10.1145/3650400.3650446

[7] Jraba, S., Elleuch, M., Ltifi, H., & Kherallah, M. (2025). Enhanced YOLOv8 Framework for Early Detection of Alzheimer's Disease Using MRI Scans. *Proceedings of the 17th International Conference on Agents and Artificial Intelligence*, 1229-1237. https://doi.org/10.5220/0013315300003890

[8] Kumar, A., & Dhanalakshmi, R. (2024). EYE-YOLO: a multi-spatial pyramid pooling and Focal-EIOU loss inspired tiny YOLOv7 for fundus eye disease detection. *International Journal of Intelligent Computing and Cybernetics*, *2024*(3), 503-522. https://doi.org/10.1108/ijicc-02-2024-0077

[9] Lin, G., Tang, Y., Zou, X., Li, J., & Xiong, J. (2019). In-field citrus detection and localisation based on RGB-D image analysis. *Biosystems Engineering*, 186:34-44. https://doi.org/10.1016/j.biosystemseng.2019.06.019

[10] Matache, M.G., Marin, F.B., Persu, C.I., Cristea, R.D., Nenciu, F., & Atanasov, A.Z. (2026). Autonomous Tomato Harvesting System Integrating AI-Controlled Robotics in Greenhouses. *Agriculture*, *16*(8), 1-28. https://doi.org/10.3390/agriculture16080847

[11] Mu, G., Zhang, H., Fei, Z., Gao, G., & Zhang, H. (2025). Defect detection method of Chinese mitten crab based on improved EfficientVit. *International Journal of Simulation and Process Modelling*, *22*(1-2), 47-59. https://doi.org/10.1504/IJSPM.2025.148289

[12] Qin, D., Leichner, C., Delakis, M., Fornoni, M., Luo, S., Yang, F., Wang, W., Banbury, C., Ye, C., Akin, B., Aggarwal, V., Zhu, T., Moro, D., & Howard, A. (2017). MobileNetV4: Universal Models for the Mobile Ecosystem. *18th European Conference*, 78-96. https://doi.org/10.1007/978-3-031-73661-2_5

[13] Rong, J., Hu, L., Zhou, H., Dai, G., Yuan, T., & Wang, P. (2024). A selective harvesting robot for cherry tomatoes: Design, development, field evaluation analysis. *Journal of Field Robotics*, *41*(8), 2564-2582. https://doi.org/10.1002/rob.22377

[14] Shao, G., Sun, Q., Gao, Y., Zhu, Q., Gao, F., & Zhang, J. (2023). Sub-Pixel Convolutional Neural Network for Image Super-Resolution Reconstruction. Electronics, 12(17), 3572. https://doi.org/10.3390/electronics12173572

[15] Wang, L. C., Chu, Y. C., Huang, Y., & Lian, F. L. (2024). Enhancement on Target-Gripper Alignment: A Tomato Harvesting Robot with Dual-Camera Image-Based Visual Servoing. 2024 IEEE International Conference on Robotics and Automation (ICRA), 13611-13617. https://doi.org/10.1109/icra57147.2024.10610772

[16] Yin, H., Sun, Q., Ren, X., Guo, J., Yang, Y., Wei, Y., Huang, B., Chai, X., & Zhong, M. (2023). Development, integration, and field evaluation of an autonomous citrus-harvesting robot. *Journal of Field Robotics*, *40*(6), 1363-1387. https://doi.org/10.1002/rob.22178

[17] Zhang, G, Li, L., Zhang, Y., Liang, J, & Chun, C. (2025). Citrus pose estimation under complex orchard environment for robotic harvesting. *European Journal of Agronomy*, *162*, 1-11. https://doi.org/10.1016/j.eja.2024.127418

[18] Zhang, Z. (2000). A Flexible New Technique for Camera Calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(11), 1330-1334. https://doi.org/10.1109/34.888718

[19] Zhu, F., Zhang, W., Li, Z., Gao, T., & Zhao, Q. (2025). A YOLOv7-based robotic harvesting system for Agaricus bisporus using a depth camera. *Turkish Journal of Agriculture & Forestry*, *49*(2), 380-396. https://doi.org/10.55730/1300-011X.3272

[20] Zhu, F., Zhang, W., Wang, S., Jiang, B., Feng, X., & Zhao, Q. (2024). Apple-Harvesting Robot Based on the YOLOv5-RACF Model. *Biomimetics*, *9*(8), 1-24. https://doi.org/10.3390/biomimetics9080495