Improved YOLOv8s with Swin Transformer and Depthwise Convolutions for Small-Target Pepper Detection and Localization in Agricultural Robotics

Zhiyuan Tan[1,2], Jianneng Chen[1,2*], Chuanyu Wu[1,2,] Leiying He[1,2], Kun Yao[1,2]
[1]College of Mechanical Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China
[2]Key Laboratory of Planting Equipment Technology of Zhejiang Province, Hangzhou 310018, China
E-mail: chenjianneng_zj@outlook.com

Keywords: Pepper;YOLOv8s model;Swin Transformer; recognition and localization; deep learning

Received: Septeber 8, 2025

A recognition and localization system for chili picking robots was developed based on an improved YOLOv8s model and a RealSense depth camera. The proposed model integrates the Swin Transformer, DW Conv, and C2 modules into the YOLOv8s framework to enhance small-target detection and reduce computational complexity. A dataset containing 2,000 field images of Chaotian pepper (Capsicum frutescens L.) was collected under varying lighting and occlusion conditions, and divided into training, validation, and test sets (7:2:1). To validate the effectiveness of the proposed approach, comparative experiments were conducted against YOLOv5, YOLOv6, YOLOv7, and the original YOLOv8s models. Ablation studies demonstrated that each added component improved model performance, with the combined integration achieving the best results. The improved YOLOv8s model reached a mean Average Precision (mAP) of 82.7%, Recall (R) of 93.0%, and Precision (P) of 79.0%, representing respective increases of 3.4%, 3.0%, and 5.7% compared with the baseline YOLOv8s. These results confirm that the improved YOLOv8s model achieves accurate and efficient chili recognition and localization suitable for robotic harvesting applications.

Povzetek: Predstavljen je izboljšan sistem za zaznavanje čilijev pri robotskem obiranju, ki z nadgrajenim modelom YOLOv8s omogoča natančnejše in učinkovitejše prepoznavanje ter lokalizacijo plodov v realnih razmerah.

1 Introduction

Pepper originated from the tropical and subtropical regions of Central and South America, belonging to the genus Capsicum of Solanaceae. The fruit is rich in capsaicin , which can be used in food, medical treatment, preservation and other fields[1-3]. Since the 14th Five-Year Plan, the annual planting area of pepper in China has risen steadily, accounting for 8%~10% of the total planting area of vegetables in China, with an output value of about 250 billion yuan. The planting area and output value rank first among vegetables[4-5]. With the gradual maturity of pepper cultivation technology, planting area steadily increased[6]. However, at present, the harvest of pepper mainly depends on manual picking, with high labor intensity and high picking cost. In some areas, pepper planting areas are harvested by industrial harvesters. Although machine picking can improve efficiency, some peppers may be damaged during picking, and the impurity rate is high. Because Chaotian pepper is edible pepper, the color and integrity of pepper after harvest are highly required, and mechanical harvesting will seriously reduce the quality of pepper. To improve the picking rate, reduce the damage and waste of pepper, it is of great significance to develop a special pepper picking robot for the characteristics of pepper. Identifying mature pepper is one of the key technologies of pepper picking robot[7-8].

A recognition and location system based on improved YOLOv8s model and realsense depth camera was constructed to solve the environmental problems of branches and leaves obscuring and fruit overlapping during pepper picking. Because peppers usually appear in images, it is particularly important to improve the detection ability of small targets. The SPPF module in the backbone network is replaced by the Swin Transformer module to improve the detection of small targets. The Conv network in YOLOv8s structure is replaced by DW Conv neural network. Finally, the C2F module in the neck is replaced by C2 module to reduce the computational complexity and reduce the number of parameters. The experimental results show that the model recall R reaches 0.93, the average precision mAP reaches 0.827, the accuracy P is 0.79. when the depth value of pepper in the camera is between 20cm and 30cm, the error between the measured value and the actual spatial coordinate of pepper in the camera coordinate system is the smallest, and the minimum error is only 1mm, which meets the recognition and positioning accuracy requirements of pepper picking robot. Liu Sixing et al. adopted YOLOv3 model and combined with RealSense depth camera to realize pepper recognition and three-dimensional spatial positioning in

different scenes[9].Wei Tianyu et al. introduced bidirectional feature pyramid network based on improved YOLOv5s model to further improve the detection accuracy of pepper under different lighting conditions, and also obtained the spatial position information of pepper by using RealSense depth camera[10].Chen Dexin et al.proposed Efficientnet's YOLOv3 deep convolutional neural network to identify and locate bell pepper fruits in view ofcomplex factors suchas light changes, branches and leaves occlusion and dense overlapping of fruitsin natural scenes[11]. In order to improve the picking efficiency and ensure the freshness of fruits and vegetables, Wang Long et al.proposed an improved CNN model FRRN-DCov-PAM and a domain adaptive semantic segmentation method to realize accurate segmentation and recognition of sweet pepper fruits and plants in greenhouse at night[12]. Li Lian et al. used convolutional neural network to recognize pepper images and achieved high recognition rate[13]. Zhong Shihao constructed an image recognition and spatial positioning scheme forclustered pepper fruitsbased on deep learning and deep vision, aiming to improve picking efficiency and reduce damage rate[14].Huang Huacheng used hyperspectral imaging technology to identify the maturity and damage of pepper in the field, which provided a new idea for pepper quality detection[15].

Although advanced computer vision techniques such as Swin Transformer and DWConv have been widely applied in object detection, few studies have integrated these methods specifically for agricultural picking tasks. This paper does not claim algorithmic novelty in the development of new architectures; rather, it focuses on the innovative integration of well-established components into the YOLOv8s framework to address the challenges of chili recognition and localization under complex occlusion conditions. By combining Swin Transformer, DWConv, and C2 modules in a unified framework, this study demonstrates how targeted structural optimization can significantly enhance detection accuracy and computational efficiency for small agricultural targets.

Table 1: Summary of existing studies on pepper recognition and localization

Author / Year	Model Used	Dataset Type / Size	mAP (%)	Recall (%)	Main Weaknesses
Liu et al. (2024)	YOLOv3 + RealSense	1500 field images	78.6	89.2	Limited small-object detection; high miss rate under occlusion
Wei et al. (2023)	Improved YOLOv5s + BIFPN	1800 greenhouse images	80.4	90	Reduced accuracy under complex lighting
Chen (2023)	EfficientNet-YOLOv3	1200 indoor images	77.5	86.3	Poor robustness in dense fruit clusters
Wang (2022)	FRRN-DCov-PAM	1000 night greenhouse images	75.2	82.7	Heavy computation; unsuitable for real-time detection

2 Materials and methods

2.1 Research gap

Ultralytics released YOLOv8 in 2023[16]providing a completely new SOTA model that includes object detection networks at P5 640 and P6 1280 resolutions and an instance segmentation model based on YOLACT. YOLOv8n/YOLOv8s/YOLOv8m/YOLOv8l/YOLOv8x models of different sizes are provided to meet the needs of different scenarios.

The backbone network and Neck part of YOLOv8 replace the C3 structure of YOLOv5 with a C2f structure with more abundant gradient flow, and adjust the number of channels for different scale models. The Head part is replaced by the current mainstream Decoupled Head structure, separating the classification and detection heads; at the same time, it is also replaced by Anchor-Free from Anchor-Based. YOLOv8 uses Task-Aligned Assigned positive and negative sample matching and introduces Distribution Focal Loss(DFL). The data enhancement part of the training introducesthe last 10 epochs in YOLOX to turn off Mosaic enhancement operations.

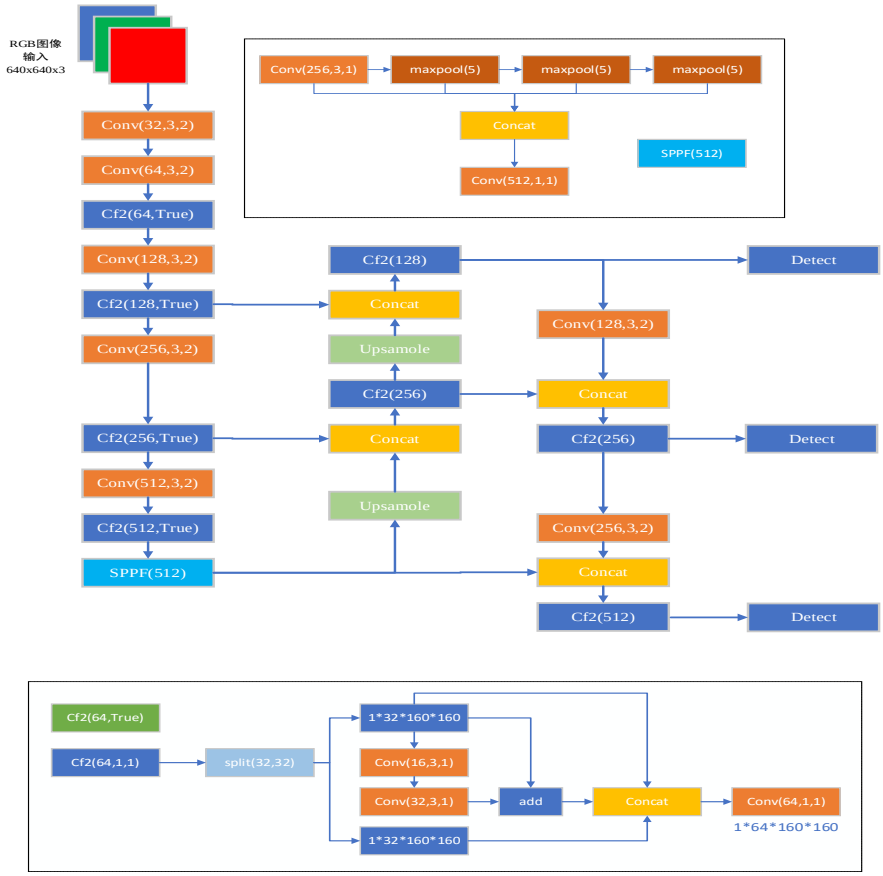

Figure 1: YOLOv8s network structure

2.2 YOLOv8 network improvements

YOLOv8s is used as the basic network of pepper detection network in this paper.SWIN transformer module is adopted to replace SPPF module in backbone network, and small target detection is improved. All Conv networks in YOLOv8s structure are replaced by DW Conv neural networks, and finally all C2F modules in the neck are replaced by C2 modules,which reduces the computational complexity and reduces the number of parameters.The improved YOLOv8s algorithm improves the average precision mAP by 3.4% and the recall by 3%.

The Swin-T (Tiny) configuration was adopted, consisting of four stages (2, 2, 6, 2 layers) with 4×4 token size and 640×640 input resolution. ImageNet-1K pretrained weights were used to accelerate convergence. The integration increased computational cost by only 4.5% while significantly improving small-object detection accuracy

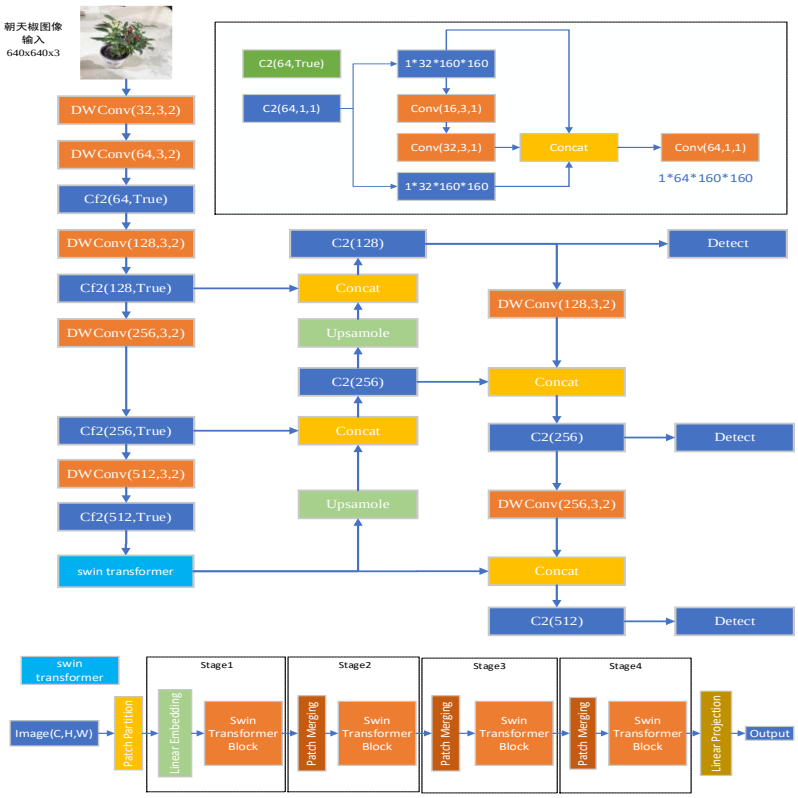

Figure 2: Improved YOLOv8s network structure

2.3 DW Conv convolution

DW Conv (Depthwise Convolution) is a lightweight convolutional operation designed to reduce computational complexity while maintaining feature extraction capability. It is a core component of Depthwise Separable Convolution (DSC), which consists of two parts: Depthwise Convolution and Pointwise Convolution (1×1 convolution)[17]. In the Depthwise Convolution stage, each channel of the input feature map is convolved independently with its own filter, capturing spatial information efficiently. In the Pointwise Convolution stage, the 1×1 filters combine the outputs of all channels to integrate cross-channel information. This decomposition significantly reduces the number of parameters and computational cost compared to standard convolution, while maintaining comparable accuracy. In this study, all standard Conv layers in the YOLOv8s model were replaced by DW Conv layers to accelerate computation and improve model efficiency without sacrificing detection performance[18].

Depthwise Separable Convolution (DSC) follows the design proposed by Howard et al. (MobileNets, 2017). DW Conv was applied throughout all convolutional layers of the YOLOv8s backbone and neck to reduce parameters and computational cost without loss of accuracy.

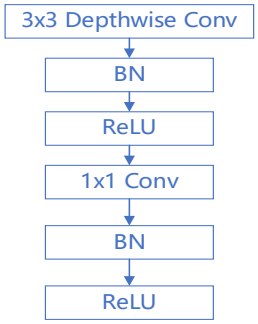

Figure 3: Structure of deep separable convolution (DSC)

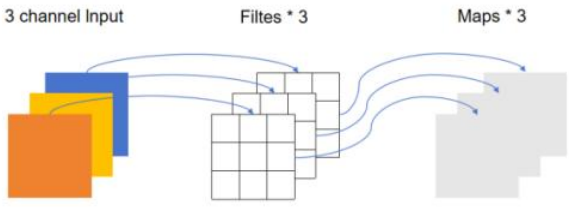

Figure 4: DW convoperation process

2.4 Swin transformer

Swin Transformer is a vision task-based Transformer model architecture that can serve as a universal backbone model for machine vision. It is composed of multiple MMBasicLayers (i.e. Stage1,2,3,4 in the figure). Each MMBasicLayer is composed of a series of Swing Transformer Blocks connected by Linear Embedding or Patch Merging. It is a hierarchical visual Transformer. When processing images, Swin Transformer uses translation strategy. This method can better capture local and global information of images, reduce information loss, and improve the processing ability of the model, so that the model can efficiently process image features of different scales[19]. By replacing the SPPF module in the backbone network with the Swin Transformer and combining the detection head of YOLOv8s, the detection effect of targets of different scales can be effectively improved. In many application scenarios, especially for the detection of small targets such as pepper, the effect is improved more significantly.

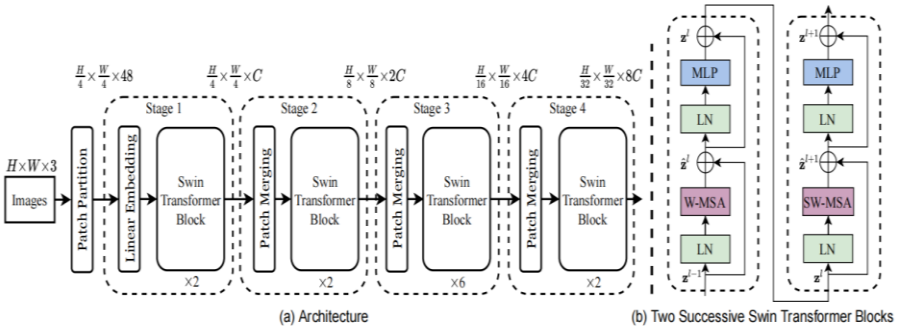

Figure 5: (a) Architecture of the swint transformer (Swin-T);(b) Two consecutive swin transformer blocks

2.5 Module C2

The C2 module usually contains only one or two convolutional layers. Its main purpose is to complete basic feature extraction tasks without significantly increasing computational complexity. Its structure is simple, and the size and number of convolution kernels are usually small. Convolutional layers are usually followed by a nonlinear activation function, such as ReLU, which is used to introduce nonlinear features[20] and has the structure shown in Figure 6.

In contrast, the C2F module inherent in YOLOv8 contains multiple convolutional layers and Bottleneck layers, while the C2 module contains only convolutional layers; the C2 module is simpler than the C2F module, and the calculation amount can be significantly reduced by using the C2 module; and the C2 module usually contains fewer parameters than the C2F module. Replacing the C2F module with the C2 module in the head of YOLOv8s simplifies the model structure, reduces computational complexity and parameter count, and improves its applicability in resource-constrained environments. The test results show that the replacement model has a significant improvement in computational efficiency, while the impact on detection performance remains within an acceptable range. Although the accuracy of the model decreases slightly in some cases, the overall detection effect can still meet the needs of practical applications.

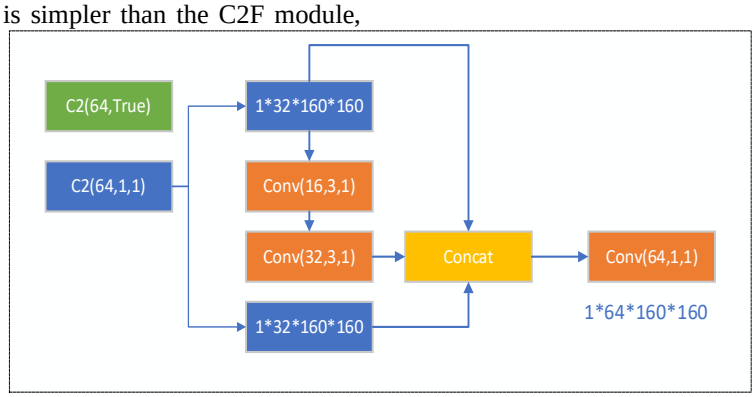

Figure 6: Structure of C2

2.6 Spatial positioning of pepper

In addition to using the improved YOLOv8s algorithm to identify and extract the two-dimensional image of pepper, it is also necessary to combine the binocular depth camera to extract and locate the three-dimensional coordinates of pepper. In this paper, RealSense D435i binocular depth camera combined with the improved YOLOv8s algorithm is selected to locate pepper. Download the SDK function pack of this camera from Intel®RealSenseTM official website, and realize data acquisition of depth camera by calling the interface provided by SDK. After obtaining the image data, the improved YOLOv8s algorithm is used to detect and recognize pepper. The algorithm outputs the detection frame of pepper in the 2D image, each detection

frame contains the center point coordinates (XY coordinates), width, height and confidence. According to the confidence threshold, reliable detection results are screened out, and the positioning point is determined as the center point of the detection frame. In order to convert into three-dimensional coordinates, depth values (Z coordinates) of corresponding positions are extracted from the depth map using depth information provided by the depth camera, and pixel coordinates and depth values are converted into actual three-dimensional spatial coordinates using a coordinate conversion method by using intrinsic and extrinsic camera parameters.

3 Data acquisition and model performance evaluation

Model training was conducted on Windows11 operating system, and various virtual environments were built for network model deployment, among which the deployment environment of YOLOv8s improved model was Python version: 3.10.14, Pytorch version: 2.3.1, Conda version: 24.4.4, CPU was i9-13980HK, GPU was NVIDIA GeForce RTX 4070.The SGD optimizer was used to update parameters with Epoch = 100, initial learning rate = 0.01, batch size = 16, momentum = 0.937, and weight decay coefficient = 0.0005.

Table 1: Experimental configuration and training parameters

Name	Configuration
Operating system	Windows11
CPU	i9-13980HK
GPU	GeForce RTX 4070
Learning framework	Pytorch 2.3.1
Acceleration environment	CUDA 11.8
Epoch	100
Batch	16

3.1 Data set

The dataset was collected at the Chaotian pepper planting base in Shandong Province during the harvest season (August to September). Images were captured under natural lighting conditions at different times of the day, including morning, noon, and afternoon, to reflect various illumination intensities. Weather conditions during data acquisition included both sunny and cloudy days, ensuring environmental diversity. The original image resolution was 1920×1080 pixels before being resized to 640×640 for model training. To enhance generalization, the dataset was divided into training, validation, and test sets that were collected under slightly different environmental and lighting conditions. This setup ensures that the improved YOLOv8s model can effectively detect pepper fruits under complex field scenarios with variable brightness, shading, and occlusion levels.

This study focuses exclusively on Chaotian pepper (commonly called bird's eye chili). The words "chilli" and "pepper" are used interchangeably to denote this same crop species (Capsicum frutescens L.).

The dataset used in this study is private and was independently collected at the Chaotian pepper plantation base. Representative statistics are as follows: image resolutions range from 1800×1080 to 1920×1080 pixels; the average pepper size is 42×28 pixels, covering 0.18%–0.35% of the image area; and approximately 27% of the samples include partial occlusion by leaves or branches. A public version with anonymized annotations will be released after publication.

Data augmentation was applied to approximately 60% of the training images to enhance model robustness. The main augmentation techniques included Mosaic (25%), random brightness and contrast adjustment (20%), and horizontal flipping (15%). Experiments showed that Mosaic augmentation contributed most significantly, improving mAP by about 2.1%. All augmentations were applied exclusively to the training set, while the validation and test sets remained unchanged for unbiased evaluation.

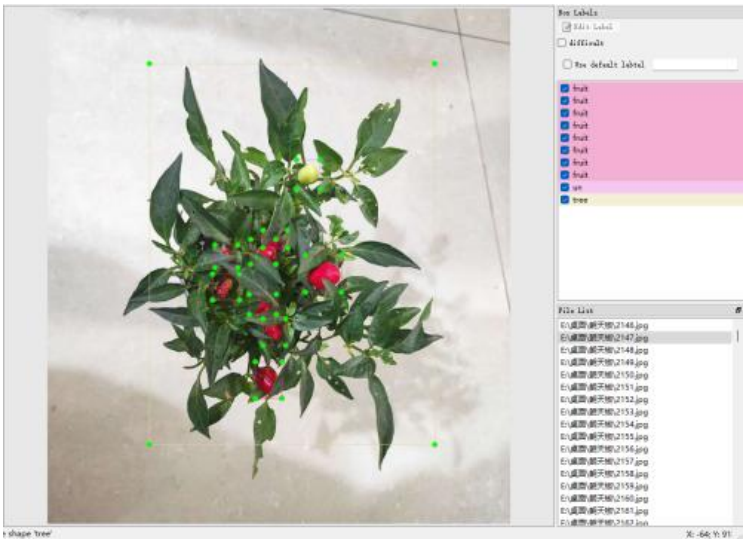

Figure 7: Data set labeling

3.2 Evaluation indicators

Precision (P/%),Recall (R/%) and the mean Average Precision (mAP) (mAP) were used as indexes to evaluate the performance of the algorithm. The expression is as follows:

$$P = \frac{TP}{TP+FP} \times 100\% \quad (1)$$

$$R = \frac{TP}{TP+FN} \times 100\% \quad (2)$$

$$mAP = \frac{1}{C}\sum_{i=1}^{C}\int_0^1 P(R)d(R) \times 100\% \quad (3)$$

$$P = \frac{TP}{TP+FP}, \quad R = \frac{TP}{TP+FN}, \quad mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \quad (4)$$

$$X = (u - c_x)\cdot Z/f_x, \quad Y = (v - c_y)\cdot Z/f_y \quad (5)$$

TP represents the correctly detected pepper target, FP represents the incorrectly detected pepper target, FN represents the actually existing but undetected pepper target, and N represents the number of label categories.

Where mAP@0.5 (IoU threshold = 0.5) was used for evaluation.The transformation matrix was derived from RealSense SDK calibration, ensuring accurate coordinate conversion between pixel and world spaces.

4 Test results and analysis

4.1 Ablation test

In order to verify the detection effect of three improved methods on pepper, ablation experiments and comparative analysis were carried out on pepper data set. The same equipment and parameter configuration was used during the test to ensure the rigor of the test. Compared with the original YOLOv8s model, the performance of the three improved algorithms is improved in different degrees, and the calculation amount is also reduced. The test results are shown in Table 2.

Table 2: YOLOv8s modified ablation test results

Trial No.	C2	DW Conv	Swin Transformer	mAP/%	Calculated quantities(GFLOPs)
1	×	×	×	79.3	28.4
2	√	×	×	82.2	28
3	√	√	×	82.3	23.4
4	√	√	√	82.7	23.1

In Table 2, “Calculated quantities (GFLOPs)” refer to the total floating-point operations required for a single forward pass of the entire model. To reflect real-world inference performance, the average inference time per image was also measured. The improved YOLOv8s achieved an inference speed of 12.4 ms/image on an NVIDIA RTX 4070 GPU, indicating high computational efficiency suitable for real-time applications.

4.2 Comparison of test results of different algorithms

In order to further evaluate the detection performance of the improved YOLOv8s algorithm for pepper, the improved algorithm is compared with YOLOv5, YOLOv6, YOLOv7 and the original YOLOv8s target detection algorithm. As can be seen from Table 3, the average precision mean and recall R of the improved YOLOv8s algorithm are better than other algorithms. Compared with the original YOLOv8s, the mean Average Precision (mAP) increases by 3.4% to 82.7%; the recall

rate increases by 3% to 93.0%.

Table 3: Test results of different algorithms in the test set

serial number	model	mAP@0.5/%	R/%	P/%
1	YOLOv5	78.2	89	75.6
2	YOLOv6	65.1	67	60.4
3	YOLOv7	67.8	70	63.1
4	YOLOv8s	79.3	90	73.3
5	YOLOv8s improvements	82.7	93	79.0

4.3 Application testing evaluation

To prove the effectiveness of the improved YOLOv8s algorithm in detecting bird's eye chili peppers, a recognition comparison was made between the original YOLOv8s algorithm and the improved algorithm.Select complex scenes with branches and leaves blocking and fruits overlapping under different circumstances. As shown in Figure8, the original algorithm has the problems of false detection and missed detection in complex scenes, and the confidence of detecting pepper is low. The improved YOLOv8s algorithm effectively reduces the false detection and missed detection problems existing in the original YOLOv8s recognition and detection process, and the confidence of identifying pepper is improved. To sum up, the improved YOLOv8s algorithm has better detection performance for pepper.

The field experiment included 300 images from 10 different scenes under varying illumination (morning, noon, and late afternoon) and weather conditions (sunny and cloudy). All images were manually annotated using LabelImg, with successful detection defined as IoU ≥ 0.5 and confidence ≥ 0.7. Across five repeated runs, the improved YOLOv8s achieved mAP = 82.7% ± 0.4 and Recall = 93.0% ± 0.5, confirming stable performance under real-world variability.

(a) Original YOLOv8s (b) Improved YOLOv8s

Figure 8: Comparison of recognition effect between original YOLOv8s and improved YOLOv8s

To further verify the effectiveness of the improved YOLOv8s algorithm in practical application scenarios, we conducted field scenario tests at the Chao Tian Pepper planting base. The results show that in the complex field environment, the recognition confidence of the improved YOLOv8s algorithm for bird's eye chili fruits has been significantly enhanced compared with the original algorithm. For some of the bird's eye chili targets that are half-shaded by leaves and have fruits stacked on top of each other, the confidence levels have all increased, as shown in Figure 9. The field test results are consistent with the test conclusions of the laboratory simulation of complex scenarios, further confirming that the improved algorithm has a more reliable detection ability for bird's eye chili in the real agricultural environment, can more accurately identify the target, and provides strong technical support for subsequent practical applications such as automated picking and yield estimation.

(a) Original YOLOv8s (b) Improved YOLOv8s

Figure 9: compares the recognition effects of the original YOLOv8s and the improved YOLOv8s in the field scene

4.4 Three-dimensional positioning accuracy

The experiment mainly selects the depth value (Z coordinate) in the depth map of pepper at different positions from the camera, and compares it with the actual spatial coordinate value of pepper in the camera coordinate system to study the influence of Z value on the ranging accuracy. In order to comprehensively evaluate the influence of Z value on ranging accuracy, several Z values were selected for experiments. Since the minimum measurement distance of RealSense D435i binocular depth camera is 15cm, Z values here are20cm, 30cm, 40cm, 50cm, 60cm, 70cm, 80cm and 90cm respectively, covering the distance range from closer to farther, which can more accurately understand the performance of ranging accuracy at different distances. In the experiment, the measurement data of pepper relative to camera coordinate system under these Z values were re corded in detail. The results show that the error increases with the Z value, and the minimum error is 1mm when Z value is between 20cm and 30cm. When Z value is between 20cm and 30cm, the distance between recognition camera and pepper can meet the positioning accuracy requirement of picking robot.

5 Discussion

Compared with YOLOv5, YOLOv6, and YOLOv7, the improved YOLOv8s model demonstrates higher mAP (82.7%) and Recall (93.0%) while reducing computational complexity from 28.4 GFLOPs to 23.1 GFLOPs. The integration of the Swin Transformer enhances contextual awareness and improves small-object detection under occlusion, while DW Conv and C2 modules significantly decrease the number of parameters and enable faster inference. These modifications achieve a balance between detection accuracy and model efficiency. Although slight precision loss occurs in some low-light cases, the model performs more robustly in natural field environments with overlapping fruits and irregular lighting. This improvement translates directly into practical value for automated field operations, enabling more accurate recognition and localization of chili targets for robotic picking systems.

6 Conclusions

In order to solve the problems of low recognition efficiency, inaccurate recognition accuracy and missed detection of occlusion target of pepper picking robot,a pepper target recognition detection algorithm based on improved YOLOv8 algorithm is proposed in this paper.By designing SWIN Transformer module to replace SPPF module in YOLOv8s backbone network,the effect of small target detection is enhanced; all Conv network in YOLOv8s structure is replaced by DW Conv neural network, and finally all C2F modules in neck are replaced by C2 modules, so as to reduce computational complexity and reduce parameter number.The experimental results show that compared with the original YOLOv8s algorithm, the mAP of the improved YOLOv8s is increased by 3.4%,the confidence is higher,and the detection performance of pepper target in occlusion environment is also improved.The detection accuracy and positioning accuracy can meet the actual picking requirements of pepper.This algorithm can provide technical support for pepper picking robot.

In this study, an improved YOLOv8s-based recognition and localization model was proposed for chili picking robots. The novelty of this work lies not in the invention of new algorithms, but in the tailored integration and empirical validation of Swin Transformer, DWConv, and C2 modules to enhance small-target detection in complex field environments. Experimental results confirm that this integrated approach achieves higher precision and recall while maintaining lightweight computation, providing a practical and effective solution for agricultural automation.

List of abbreviations

YOLO: You Only Look Once
P: Precision
R: Recall
mAP: Mean Average Precision

DFL: Distribution Focal Loss
DW Conv: Depthwise Convolution
DSC: Depth Separable Convolution
PW Conv: Pointwise Convolution
EW Conv: Expanded Width Convolution
SPPF: Spatial Pyramid Pooling Fast
ReLU: Rectified Linear Unit
SDK: Software Development Kit
TP: True Positive
FP: False Positive
FN: False Negative
TN: True Negative

Funding

This research is supported by the Key Project of National Natural Science Regional Innovation Joint Fund (No.U23A20175).

Authors' contributions

Materials preparation and data analysis were completed by Z.Y.T. J.N.C. was responsible for supervision and project funding. C.Y.W. provided test site and facility support. Z.Y.T. and K.Y. designed the overall framework. L.Y.H. developed the test protocols. Z.Y.T. and K.Y. conducted the field tests. The first draft of the manuscript was written by Z.Y.T., and all authors commented on, read, and approved the final version.

References

[1] Zou Xuexiao, Ma YQ, Dai XZ, et al. Pepper dissemination and industrial development in China [J]. *Acta Horticultura Sinica*, 2020, 47 (09):1715-1726.

[2] Zou Xuexiao, Zhu Fan. Origin, evolution and cultivation history of pepper [J]. *Acta Horticultura Sinica*, 2022, 49 (06):1371-1381.

[3] Zou Xuexiao, Hu Bo Wen, Xiong Cheng, et al. Review and prospect of pepper breeding in China in the past 60 years [J]. *Acta Horticultura Sinica*, 2022, 49 (10):2099-2118.

[4] Saddik A, Latif R, Taher F, El Ouardi A, Elhoseny M. Mapping agricultural soil in greenhouse using an autonomous low-cost robot and precise monitoring. *Sustainability.* 2022 Dec;14(23):15539. doi:10.3390/su142315539.

[5] Zuo MHQ, Zhao YH, Yu SS. Industrial robot applications and individual migration decision: evidence from households in China. *Humanities & Social Sciences Communications.* 2024 Aug 9;11(1):1022. doi:10.1057/s41599-024-03542-z.

[6] Yu KZ, Shi Y, Feng JH. The influence of robot applications on rural labor transfer. *Humanities & Social Sciences Communications.* 2024 Jun 20;11(1):796. doi:10.1057/s41599-024-03333-6.

[7] Aivazidou E, Tsolakis N. Transitioning towards human-robot synergy in agriculture: a system thinking perspective. *Systems Research and Behavioral Science.* 2023 May;40(3):536–551. doi:10.1002/sres.2887.

[8] Adamides G, Katsanos C, Parmet Y, Christou G, Xenos M, Hadzilacos T, Edan Y. HRI usability evaluation of interaction modes for a teleoperated agricultural robotic sprayer. *Applied Ergonomics.* 2017 Jul; 62:237–246. doi: 10.1016/j.apergo.2017.03.008.

[9] Liu Sixing, Li Shuang, Miao Hong, et al. Research on identification and localization of pepper picking robot based on YOLOv3 in different scenes [J]. *Agricultural Mechanization Research*, 2024, 46 (02):38-43.

[10] Wei Tianyu, Liu Tianhong, Zhang Shanwen, et al. Identification and localization method of pepper picking robot based on improved YOLOv5s [J]. *Journal of Yangzhou University (Natural Science Edition)*, 2023, 26 (01):61-69.

[11] Chen Dexin. Fruit recognition and location of bell pepper based on binocular vision [D]. Henan Agricultural University, 2023.

[12] Wang Long. Semantic segmentation algorithm based on convolutional neural network and its application in sweet pepper image recognition [D]. Jiangsu University, 2022.

[13] Li Lian, Ding Wenkuan. Pepper recognition based on convolutional neural network [J]. *Journal of Tianjin University of Technology*, 2017, 33 (03):12-15.

[14] Zhong Shihao. Research on clustering pepper target recognition and localization algorithm based on deep learning [D]. Guizhou Normal University, 2024.

[15] Huang Huacheng. Study on maturity and damage identification of fresh pepper based on hyperspectral technology [D]. Guizhou University, 2023.

[16] Terven J, Córdova-Esparza D M, Romero-González J A. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS[J]. *Machine Learning and Knowledge Extraction*, 2023, 5(4): 1680-1716.

[17] Guo, Yunhui, et al. "Depthwise convolution is all you need for learning multiple visual domains. " Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. No. 01. 2019.

[18] Chollet, François. "Xception: Deep learning with depthwise separable convolutions. " Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[19] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows. " Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[20] Faxon HO. Small farmers, big tech: agrarian commerce and knowledge on Myanmar Facebook. *Agriculture and Human Values.* 2023 Sep;40(3):897–911. doi:10.1007/s10460-023-10446-2.