International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 14, Issue 8 [August 2025] PP: 45-52

Research on Recognition Model of Mature Tomato Fruits Based on Improved YOLOv8n

Meng Wang, Yu Wang*, Jiahui Jin
School of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo, 454000, China
**Corresponding Author*

ABSTRACT:
Tomato fruit detection, as a key technology in intelligent picking, faces significant challenges in complex environments characterized by background interference, foliage occlusion, fruit overlapping, and the resulting scale variations. To address these issues, this paper proposes a tomato target detection algorithm based on the improved YOLOv8n. Firstly, the CBAM (Convolutional Block Attention Module) attention mechanism is embedded at the end of the YOLOv8n backbone network, which enhances the feature expression capability through the synergistic effect of channel attention and spatial attention. Secondly, the Wise-IoU loss function is introduced to dynamically adjust the weight of localization loss, thereby optimizing the model's bounding box localization capability. Experimental results show that the precision, recall, and mAP values of the improved model reach 88.1%, 87.3%, and 90.7% respectively, which are 1.3, 7.5, and 3.8 percentage points higher than those of the original YOLOv8 model, respectively.

Date of Submission: 01-08-2025 Date of acceptance: 11-08-2025

I. INTRODUCTION

As a crop with high economic value in China, tomato relies on efficient and accurate object detection technology to achieve automated detection, which is of great significance for subsequent yield prediction, quality monitoring, and automated harvesting [1]. However, in real agricultural scenarios, the diversity of tomato growth environments significantly increases the difficulty of object detection. For example, the instability of illumination, including changes in light intensity and angle, can significantly alter the appearance characteristics of tomatoes, such as color, brightness, and texture, leading to detection deviations of the detection model under different illumination conditions. Tomatoes are dense crops, and targets are prone to mutual occlusion, resulting in incomplete presentation of tomatoes in images, making it difficult to extract target features and increasing the risk of false detection and missed detection. The background of real agricultural scenarios is complex, and the shape and color of tomatoes are easily confused with the surrounding environment, making it difficult to distinguish between background and targets. The size of individual tomatoes also varies due to differences in growth stages, varieties, and planting density. All these factors increase the difficulty of tomato detection. Traditional object detection methods, such as graph segmentation [2] and color segmentation [3], have problems such as low detection accuracy, poor robustness, large computational load, and slow speed, which are difficult to meet the actual application requirements for tomato detection in complex environments.

In recent years, with the development of computer vision technology, object detection methods have undergone a transformation from handcrafted features to deep learning. Traditional object detection methods are based on information such as the size, color, and texture features of objects [4], and use methods such as random forests, support vector machines (SVM) [5], and K-means clustering algorithms [6] for fruit detection. For instance, Liu et al. [7] proposed an algorithm for automatic tomato detection in conventional color images. When using HOG and SVM for classification, it has limitations in failing to make full use of high-dimensional information and poor robustness. Since HOG features are hand-designed, they cannot automatically learn deeper image features, resulting in the model being very sensitive to illumination changes and occlusion interference, and performing poorly in complex environments.

In the field of image and video processing, the progress of deep convolutional neural networks (DCNN) has promoted the wide application of deep learning-based object detection technology in fruit and vegetable detection [8]. Compared with traditional algorithms, the greatest advantage of deep learning methods is their ability to automatically learn and extract deep-level, highly robust features from massive data, thereby achieving higher detection accuracy. This has led to a diversified development trend in current deep learning-based tomato fruit detection technology.

For example, by introducing a region proposal network (RPN), candidate target regions are generated in the image, and combined with classifiers and regressors to achieve object detection tasks. Chen Jiuhao et al.

[9] designed an improved Faster R-CNN model algorithm with ResNet-50 as the backbone network to realize the detection of bitter gourd leaf diseases under natural environmental conditions. The experimental results showed that the accuracy of the improved model was 7.54 percentage points higher than that of the original model.

YOLO, as a one-stage object detection algorithm, has received widespread attention since its proposal in 2015. On this basis, many researchers have also committed to improving the YOLO model, focusing on model optimization. For example, Huang et al. [10] proposed a YOLO algorithm with dense connections (to enhance feature utilization) and spatial pyramid pooling (to capture multi-scale information), namely DCSPP-YOLO, which improved the object detection accuracy of YOLOv2. Liu Fang et al. [11] proposed an IMS-YOLO algorithm to realize the rapid and accurate detection of tomato fruits by agricultural picking robots in greenhouse environments. Zhang et al. [12] made two improvements to YOLOv4: first, combining the GhostNet feature extraction network with the coordinate attention module; second, reconstructing the neck and head parts through depth-wise separable convolution. Li Changlu [13] optimized for small targets in the SSD algorithm by introducing the CBAM attention mechanism and feature fusion to improve the detection speed and recognition accuracy of apples. Wang Yong et al. [14] replaced the traditional Conv module with the ODConv convolution module to enhance the ability of the YOLOv5 algorithm to extract fine-grained features of apples, thereby improving the apple recognition ability in natural environments. Miao Ronghui et al. [15] introduced MobileNetv3 into the YOLOv7 model to replace the backbone feature extraction network, reducing the number of network parameters to achieve lightweight, and at the same time introducing the global attention mechanism (GAM) module to improve the feature expression ability of the network and further enhance the network detection accuracy.

II. CONSTRUCTION OF TOMATO FRUIT IMAGE DATASET

DATASET

The LaboroTomato dataset [16] is a public image collection specifically designed for object detection and instance segmentation tasks, covering the growth process of tomatoes at different maturity stages. This dataset includes images of regular (large) tomatoes and cherry (small) tomatoes, all captured in greenhouse environments. The dataset categorizes tomatoes into three main maturity stages: unripe, semi-ripe, and ripe. Each tomato object is annotated with bounding boxes for object detection tasks and contains vertices representing tomato masks for instance segmentation tasks. Additionally, the dataset provides classification information of tomatoes, distinguished by maturity and size. Maturity classification is determined by the red proportion of tomatoes: fully ripe tomatoes have a red proportion of over 90%, semi-ripe ones range from 30% to 89%, and unripe ones are between 0% and 30%. Moreover, the dataset includes two different tomato types: regular (large) tomatoes and cherry (small) tomatoes. Experts classify tomatoes based on criteria such as tomato type and red proportion.

The entire dataset contains 804 images, divided into a training set (643 images), a validation set (81 images), and a test set (81 images) in an 8:1:1 ratio. To ensure the diversity of the dataset and its relevance to practical applications, the images were captured using two different cameras, each with unique resolution and image quality. The use of different cameras introduces variations in image features, such as differences in color reproduction and clarity, which may pose challenges to the performance of computer vision models, thereby better simulating real-world scenarios. As shown in Figure 1.

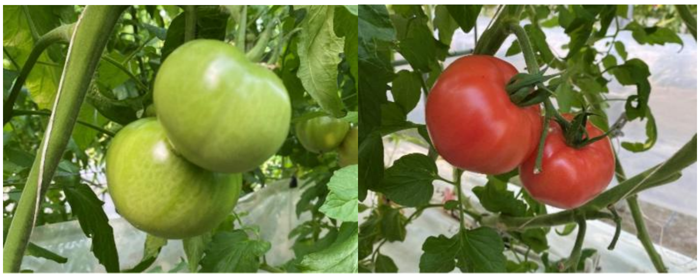
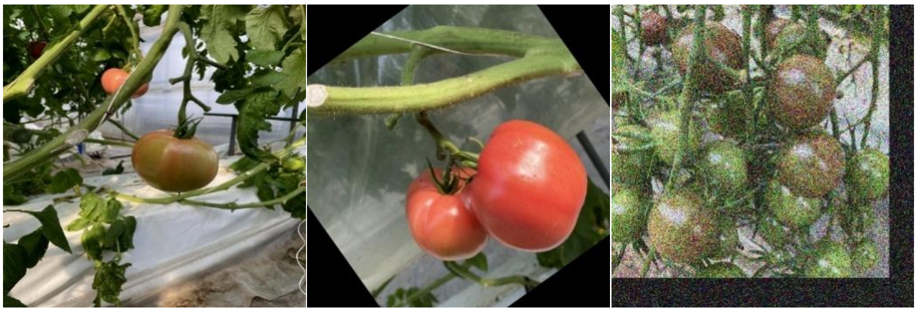


Fig.1 Partial sample data

DATA AUGMENTATION

To significantly enhance the model's generalization ability and robustness, enabling it to learn more diverse features, some images with stable quality were selected from the original images for a series of data augmentation operations. These operations include random cropping, scaling, flipping, and rotation, aiming to simulate different shooting angles and conditions, thereby enriching the model's training samples [17]. Since green tomatoes are similar in color to green branches and leaves in the greenhouse environment, this may lead to low recognition rates or even missed detections by the target detection model. Therefore, in this study, the proportion of green tomato images in the dataset was increased to improve the model's recognition capability for green tomatoes. After augmentation and screening, a dataset containing 1768 images was finally formed.

The dataset was divided into two subsets: 80% of the data was used as the training set, and 20% as the test set. According to this allocation ratio, 1415 images were assigned to the training set, whose main purpose is to support the model's main training process, ensuring that the model can fully learn the key features for tomato fruit recognition. The test set consists of 353 images, which is used to finally test the practical application effect of the model. The augmented dataset is shown in Figure 2.



ZoomrotateTranslation plus noise

Fig.2 Dataset expansion

III. TOMATO RECOGNITION METHOD BASED ON IMPROVED YOLOV8

YOLO MODEL

YOLOv8, the latest version in the YOLO series released by Ultralytics, is mainly used in fields such as object detection, image classification, and instance segmentation. Currently, it is widely applied in scenarios like autonomous driving [18], industrial automation [19], and medical image detection. Based on the model architecture of YOLOv5, YOLOv8 introduces a new detection head and sample allocation strategy to improve the model's performance and accuracy. In the Backbone section, YOLOv8 retains the SPPF module, a type of spatial pyramid pooling module that can extract multi-scale features and enhance the model's robustness. Meanwhile, to further optimize the model's performance and computational efficiency, YOLOv8 replaces the C3 module in YOLOv5 with the C2f module. Compared with the C3 module, the C2f module is more lightweight, which can reduce the model's computational complexity and memory usage while maintaining its performance, making the model more suitable for operation on resource-constrained devices. In the Head section, a decoupled head is used instead of the coupled head in YOLOv5, which separates the classification and detection heads to improve the overall performance, flexibility, and efficiency of the model. Its structure is shown in Figure 3.

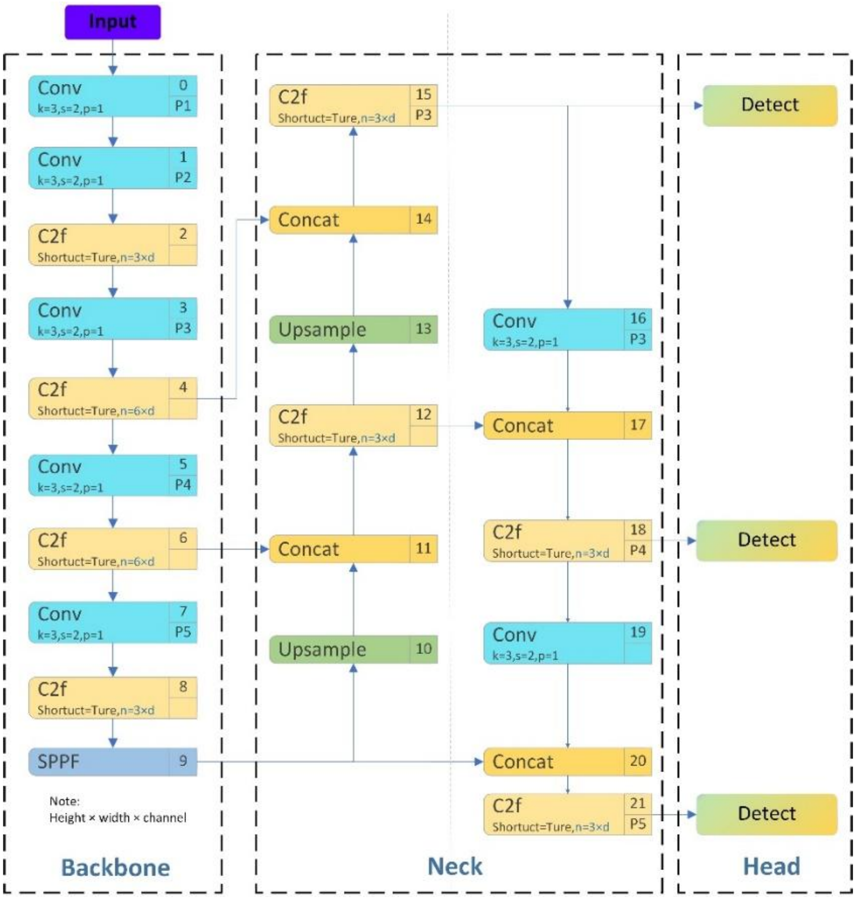


Fig.3 YOLOv8n network structure

IMPROVEMENT SCHEME FOR YOLOV8

CBAM (Convolutional Block Attention Module) is a lightweight attention module designed to enhance the feature representation capability of convolutional neural networks (CNNs). By strengthening key features and suppressing irrelevant information, CBAM improves the model's sensitivity to important visual cues, thereby achieving the goal of optimizing model performance [20]. As a lightweight attention module, CBAM can be seamlessly integrated into various existing CNN architectures (such as ResNet and VGG) without causing a significant increase in computational burden. The working principle of CBAM is shown in Figure 4, and its core mechanism is based on two sequential attention sub-modules: the Channel Attention Module and the Spatial Attention Module. These two sub-modules sequentially perform weighting processing on feature maps to highlight the important features of the image.

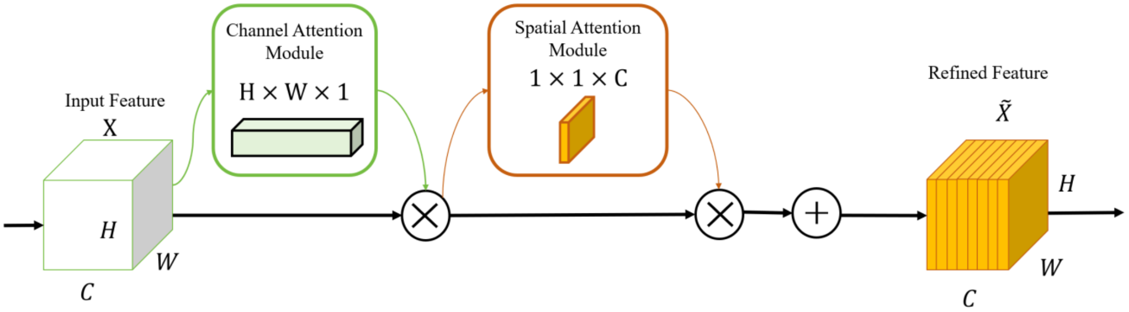


Fig.4 The architecture of CBAM

WISE-IOU LOSS FUNCTION

The built-in bounding box loss function of YOLOv8n is CIoU, which is used to measure the degree of overlap between the predicted box and the ground truth box. CIoU considers three factors: overlap area, center point distance, and aspect ratio. However, when the predicted box and the ground truth box do not intersect, CIoU cannot accurately reflect their degree of overlap, leading to inaccurate loss estimation and affecting the model optimization effect. The Wise-IoU (WIoU) loss function [21] introduces outliers based on the IoU loss function to more accurately evaluate the overlap between the predicted bounding box and the ground truth bounding box, thereby improving the detection performance of the model. The WIoU loss function is defined as follows:

$\beta = \frac{L^*_{\mathrm{IoU}}}{L_{\mathrm{IoU}}}(1)$

$L_{\mathrm{WIoU}} = \gamma L_{\mathrm{WIoUvl}} \qquad (2)$

$\frac{\partial A_{ct}}{\partial L} = W \qquad (3)$

equations (1) to (3) are for outlier degree calculation, where LIoU is used to measure the degree of overlap between the predicted box and the ground truth box; L̄IoU refers to the exponential moving average of LIoU, which is used to smooth changes in IoU loss. A non-monotonic focusing coefficient is constructed, with δ being a constant. By applying the non-monotonic focusing coefficient, LWIoUv1 is the loss function of WIoU v1.

The loss function of WIoU v3 achieves dynamic adjustment of loss weights by applying the non-monotonic focusing coefficient r to the loss function of WIoUv1, thereby optimizing the model's training process. Here, the function formula of WIoUv1 is explained as follows:

$L_{WIoU} = R_{WIoU}L_{IoU}(4)$

$R_{WIoU} = exp\,(\frac{(x-x_{gt})^2+(y-y_{gt})^2}{(W_g^2+H_g^2)^*}) \qquad (5)$

in the formula: (x, y) are the coordinates of the center point of the predicted box; (xgt, ygt) are the coordinates of the center point of the ground truth box; Wg and Hg are the width and height of the minimum enclosing box; the superscript ∗ indicates that Wg and Hg are separated from the computation graph to prevent gradients that hinder convergence.

The key advantage of the WIoU loss function lies in alleviating the excessive penalty of geometric factors (such as distance, aspect ratio, etc.) by introducing a distance attention mechanism. It dynamically adjusts the degree of attention to the distance between anchor boxes and target boxes, enhancing the model's adaptability in multi-scale targets and complex scenarios. Compared with traditional IoU loss functions, WIoU optimizes the loss function through a non-monotonic focusing mechanism, making the model more accurate when dealing with different target scales, thereby improving the bounding box regression performance and the final detection results.

IMPROVED YOLOV8 MODEL

Although the YOLOv8n algorithm has high detection accuracy and speed, for tomato targets in complex scenarios, due to factors such as variable scales and morphologies as well as susceptibility to occlusion, the model tends to ignore information of occluded objects and small targets when extracting features, often resulting in missed detections and false detections. To address the above issues, this paper improves the model from two aspects:

(1) Introduce the CBAM attention mechanism multiple times in the neck network. This enhances feature hierarchy, highlights important features, suppresses irrelevant background information, and optimizes the feature fusion process to improve detection accuracy.

(2) Replace the traditional loss function in YOLOv8n with the WIoU loss function to better balance the localization errors of mature fruit targets of different sizes, optimize bounding box localization accuracy, reduce the phenomenon of multiple box overlapping, improve the effect of non-maximum suppression (NMS), and enhance the model's adaptability in complex detection scenarios.

IV. RESULTS AND ANALYSIS

MODEL TRAINING AND EVALUATION

The experimental design for the proposed algorithm mainly adopts the PyCharm integrated development environment and a Python-based development framework. The configuration of the experimental environment in this paper is shown in Table 1. The model training undergoes a total of 300 epochs of iteration, and the optimizer used is SGD. The hyperparameters are as follows: the batch size is 32, the momentum factor is 0.937, the initial learning rate is 0.01, and the weight decay coefficient is 0.0005.

Tab1 Experiment environment

Name	Configuration
Operating System	Windows 11
CPU	AMD Ryzen 7 9700X
GPU	NVIDIA GeForce RTX 3080
Learning Framework	Pytorch 2.7.1
Acceleration Environment	CUDA 11.8
Programming Language	Python 3.9

EVALUATION METRICS

The main evaluation metrics selected in this paper are: precision (P), recall (R), mean average precision (mAP), and FLOPs. Among them, the mAP value at an IoU threshold of 0.5 is used for evaluation, i.e., mAP@0.5.

Precision (P) refers to the proportion of actually positive samples among all samples predicted as positive, which is used to measure the probability that positive samples are correctly predicted. Recall (R) refers to the proportion of actually positive samples that are predicted as positive among all actually positive samples, which is used to measure the situation of missed detections. The formulas are as follows:

$P = \frac{TP}{TP+FP}$(6)

$R = \frac{TP}{TP+FN}$ (7)

in the formulas, P stands for precision and R stands for recall. TP (True Positive) refers to samples correctly classified as positive samples; FP (False Positive) refers to samples incorrectly classified as positive samples; FN (False Negative) refers to samples incorrectly classified as negative samples.

$P_{AP} = \int_0^1 P(R)dR(8)$

$P_{mAP} = \frac{1}{n}\sum_{i=1}^{n} P_{AP}$ (9)

Computational load (GFLOPs) and number of parameters (Params) are used to evaluate the complexity of the model. The formulas are as follows:

$C = C_{in} \times C_{out}(10)$

$GFLOPS = W \times H \times K \times K \times C$ (11)

$Params = C \times K \times K(12)$

in the formulas, P stands for precision, R stands for recall, AP is the area enclosed by P and R, and mAP is the mean of APs across all detection categories.

RESULTS AND ANALYSIS

To objectively evaluate the performance of the improved model, this paper conducts experimental comparisons between the improved model and the YOLOv3, YOLOv5s, YOLOv8s, and YOLOv9s models on the tomato dataset, with the experimental results shown in Table 2.

Tab 2 Model comparison experiment

模型 Model	Precision/%	Recall/%	mAP50d/%	FLOPs/G
Fast R-CNN	72.6	67.7	77.9	955.1
YOLO v5s	83.6	81.2	85.6	16.2
YOLO v7-Tiny	81.5	81.2	84.6	56.2
YOLO v8n	86.8	79.8	86.9	8.0
Ours	88.1	87.3	90.7	7.8

In terms of precision and recall, the improved model achieves a precision of 88.1%, with detection accuracy higher than most models; the model's recall rate is 87.3%, and its target capture capability is significantly superior to other models. In terms of average precision, the improved model's mAP reaches 90.7%, which is generally higher than that of other models, demonstrating the robustness and stability of the improved model in tomato detection.

In summary, the improved model achieves the highest average detection accuracy while maintaining low resource consumption, thus showing significant advantages in the application of tomato quality detection.

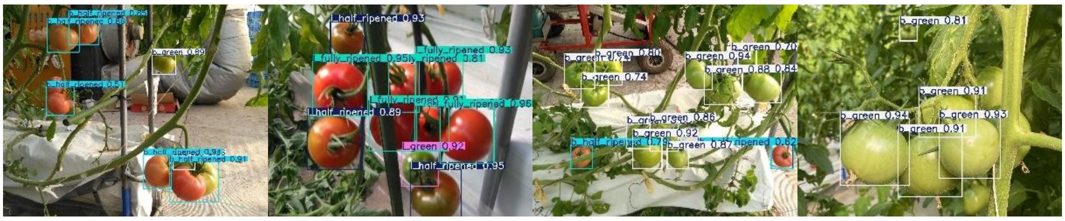


Fig.2 Detection results

To more intuitively evaluate the performance of the improved model, this paper presents the results of tomato quality detection. As shown in Figure 6, the improved model accurately classifies tomato quality, and the bounding boxes precisely locate tomato targets. Through the visualization results, it can be clearly seen that the improved model still exhibits good performance when handling detection in complex scenarios, indicating its reliability and practicality in practical applications.

V. CONCLUSION

In terms of tomato fruit recognition, the author improved the YOLOv8 model, which achieves a good balance between detection accuracy and inference speed, meeting the real-time detection requirements of tomato fruits in agricultural environments. Through practical experiments, the following conclusions are drawn:The CBAM attention mechanism is added to the Head layer of the YOLOv8 model. The improved model has better detection performance for blurred small tomato targets with complex backgrounds. Meanwhile, WIoU improves the overlap degree between the predicted boxes and the ground truth boxes by optimizing bounding box regression, reduces false detections, and makes detection more accurate.Compared with other classical object detection models, the improved YOLOv8 model achieves precision, recall, and mAP values of 88.1%, 87.3%, and 90.7% respectively; compared with the original YOLOv8 model, these values have increased by 1.3, 7.5, and 3.8 percentage points respectively.In summary, the application of the improved YOLOv8 model not only ensures the accuracy of the recognition process but also improves the accuracy and efficiency of real-time counting. Future research work can further integrate more advanced attention mechanisms, optimize convolution modules, and continuously perform model optimization on this basis to enhance the model's feature extraction capability in complex scenarios and its attention to mature tomato fruits.

Funding: This research is supported by the Cultivation Program for High-Level Research Projects under the "Double First-Class" Construction Project (grant number GCCYJ202414).

REFRENCES

[1]. C QING, Y CHENGKAI, G ZILIANG, et al. Current status and future development of the key technologies for apple picking robots. Transactions of the Chinese Society of Agricultural Engineering, 2023,39(04): 1-15.

[2]. SUN S, JIANG M, HE D, et al. Recognition of green apples in an orchard environment by combining the GrabCut model and Ncut algorithm[J]. BIOSYSTEMS ENGINEERING, 2019,187: 201-213.

[3]. HAYASHI S, YAMAMOTO S, SAITO S, et al. Field Operation of a Movable Strawberry-harvesting Robot using a Travel Platform[J]. JARQ-JAPAN AGRICULTURAL RESEARCH QUARTERLY, 2014,48(3): 307-316.

[4]. W DANYANG, L WEIHONG, L YUPING, et al. Cassava leaf disease image recognition method for imbalanced data. Journal of Chinese Agricultural Mechanization, 2025,46(03): 101-107. https://doi.org/10.13733/j.jcam.issn.2095-5553.2025.03.016

[5]. MOALLEM P, SERAJODDIN A, POURGHASSEM H. Computer vision-based apple grading for golden delicious apples based on surface features[J]. Information Processing in Agriculture, 2017,4(1): 33-40.

[6]. LUO L, TANG Y, LU Q, et al. A vision methodology for harvesting robot to detect cutting points on peduncles of double overlapping grape clusters in a vineyard[J]. COMPUTERS IN INDUSTRY, 2018,99: 130-139.

[7]. LIU G, MAO S, KIM J H. A Mature-Tomato Detection Algorithm Using Machine Learning and Color Analysis[J]. SENSORS, 2019,19(9).

[8]. SULTANA F, SUFIAN A, DUTTA P. Evolution of Image Segmentation using Deep Convolutional Neural Network : A Survey[J]. KNOWLEDGE-BASED SYSTEMS, 2020,201.

[9]. L JIUHAO, L LEJIAN, T KAI, et al. Detection of leaf diseases of balsam pear in the field based on improved Faster R-CNN. 2020,36(12): 179-185.

[10]. HUANG Z, WANG J, FU X, et al. DC-SPP-YOLO: Dense connection and spatial pyramid pooling based YOLO for object

detection[J]. Information Sciences, 2020,522: 241-258.

[11]. L FANG, L YUKUN, L SEN, et al. Fast Recognition Method for Tomatoes under Complex Environments Based on Improved YOLO. Transactions of the Chinese Society for Agricultural Machinery, 2020,51(06): 229-237.

[12]. ZHANG C, KANG F, WANG Y. An Improved Apple Object Detection Method Based on Lightweight YOLOv4 in Complex Backgrounds[J]. REMOTE SENSING, 2022,14(17).

[13]. C LI. Research on Recognition and Location of Apple Picking Robot Based on Machine Vision[D]. Lanzhou: Lanzhou University of Technology, 2022.

[14]. W YONG, T ZHAOSHENG, S XINYU, et al. Apple target detection method with different ripeness based on improved YOLOv5s. Journal of Nanjing Agricultural University, 2024,47(03): 602-611.

[15]. M RONGHUI, L ZHIWEI, W JINLONG. Lightweight Maturity Detection of Cherry Tomato Based on Improved YOLO v7. Transactions of the Chinese Society for Agricultural Machinery, 2023,54(10): 225-233.

[16]. LI R, JI Z, HU S, et al. Tomato Maturity Recognition Model Based on Improved YOLOv5 in Greenhouse[J]. AGRONOMY-BASEL, 2023,13(2).

[17]. MA J, HU C, ZHOU P, et al. Review of Image Augmentation Used in Deep Learning-Based Material Microscopic Image Segmentation[J]. Applied Sciences, 2023,13(11).

[18]. L YU-QING, Y XI-QING, C HUI-YONG, et al. Design and implementation of control system for autonomous cruise unmanned ship. Manufacturing Automation, 2022,44(10): 127-131.

[19]. W XIANG-XIANG, T LI-BIN, X XIANG-RONG. Surface defect detection method of nitrile gloves based on YOLOv5s. Manufacturing Automation, 2023,45(09): 1-4.

[20]. W MEIHUA, W ZHENXIN, Z ZUGUANG. Fine-grained Identification Research of Crop Pests and Diseases Based on Improved CBAM via Attention. Transactions of the Chinese Society for Agricultural Machinery
, 2021,52(04): 239-247.

[21]. XIONG C, ZAYED T, ABDELKADER E M. A novel YOLOv8-GAM-Wise-IoU model for automated detection of bridge surface cracks[J]. Construction and Building Materials, 2024,414: 135025.