

Article

A Lightweight Citrus Object Detection Method in Complex Environments

Qiurong Lv ^{1,2}, Fuchun Sun ^{1,2,*}, Yuechao Bian ¹, Haorong Wu ³, Xiaoxiao Li ^{1,4}, Xin Li ⁵ and Jie Zhou ¹

¹ School of Mechanical Engineering, Chengdu University, Chengdu 610106, China

² Entrepreneurship College, Chengdu University, Chengdu 610106, China

³ School of Electronic Information and Electrical Engineering, Chengdu University, Chengdu 610106, China

⁴ Institute for Advanced Study, Chengdu University, Chengdu 610106, China

⁵ Shengzhong Water Conservancy Project Operation and Management Center of Sichuan Province, Nanchong 623300, China

* Correspondence: sfc@cdu.edu.cn; Tel.: +86-199-8040-9129

Abstract: Aiming at the limitations of current citrus detection methods in complex orchard environments, especially the problems of poor model adaptability and high computational complexity under different lighting, multiple occlusions, and dense fruit conditions, this study proposes an improved citrus detection model, YOLO-PBGM, based on You Only Look Once v7 (YOLOv7). First, to tackle the large size of the YOLOv7 network model and its deployment challenges, the PC-ELAN module is constructed by introducing Partial Convolution (PConv) for lightweight improvement, which reduces the model's demand for computing resources and parameters. At the same time, the Bi-Former attention module is embedded to enhance the perception and processing of citrus fruit information. Secondly, a lightweight neck network is constructed using Grouped Shuffle Convolution (GSConv) to simplify computational complexity. Finally, the minimum-point-distance-based IoU (MPDIoU) loss function is utilized to optimize the boundary return mechanism, which speeds up model convergence and reduces the redundancy of bounding box regression. Experimental results indicate that for the citrus dataset collected in a natural environment, the improved model reduces Params and GFLOPs by 15.4% and 23.7%, respectively, while improving precision, recall, and mAP by 0.3%, 4%, and 3.5%, respectively, thereby outperforming other detection networks. Additionally, an analysis of citrus object detection under varying lighting and occlusion conditions reveals that the YOLO-PBGM network model demonstrates good adaptability, effectively coping with variations in lighting and occlusions while exhibiting high robustness. This model can provide a technical reference for uncrewed intelligent picking of citrus.

Keywords: citrus; attention mechanism; machine vision; YOLOv7; deep learning



Academic Editor: Signe Marie Jensen

Received: 21 March 2025

Revised: 8 May 2025

Accepted: 9 May 2025

Published: 12 May 2025

Citation: Lv, Q.; Sun, F.; Bian, Y.; Wu, H.; Li, X.; Li, X.; Zhou, J. A Lightweight Citrus Object Detection Method in Complex Environments. *Agriculture* **2025**, *15*, 1046. <https://doi.org/10.3390/agriculture15101046>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In global citrus production, China, Pakistan, and the United States are important citrus producers [1–3]. Among them, China's citrus farming industry is developing rapidly, with its acreage and output ranking first in the world for fruit [4]. As of 2024, China's annual citrus production is projected to be 64 million tons, with an output value of CNY 210 billion [5]. However, the harvesting of citrus has primarily been performed manually, which is labor intensive. Due to the short maturation period of citrus, the harvesting work needs to be completed within a brief time frame [6]. Therefore, the development of intelligent harvesting robots presents significant prospects [7]. In realizing intelligent harvesting, overcoming external influences, such as weather and obstructions, and quickly and efficiently

identifying citrus fruits has become a top priority. Traditional image-processing-based object detection algorithms include image feature extraction [8], color space conversion [9], and texture segmentation [10], which often struggle under real-world conditions. In actual detection scenarios, the target is frequently affected by light and obstructions, making traditional object detection methods limited in terms of accuracy, resulting in a high false detection rate and poor real-time detection results.

With the modernization and intelligent development of agriculture, deep learning networks are widely employed in agricultural image processing, demonstrating higher generalization ability and detection performance [11,12]. According to the processing flow, a representative model of two-stage detection based on deep learning is the Region-based Convolutional Neural Networks (R-CNN) series algorithms [13–15], which include two stages. The first stage extracts regions of interest (ROIs) from the input image, and the second stage classifies each ROI and regresses the bounding box. Mai et al. [16] achieved accurate detection of small fruits. Fu et al. [17] addressed the task of apple recognition under multiple categories of occlusion and significantly improved the detection accuracy of apples by using deep features to filter background objects in Faster R-CNN. Lu et al. [18] built upon the Mask-RCNN model and enhanced the multi-layer feature representation of citrus targets by integrating the CB-Net network, achieving good results in the detection of green citrus fruits. Although the two-stage detection model represented by the R-CNN series of algorithms has achieved excellent detection results, its network structures are complex and computationally intensive. The most famous and widely used single-stage detection models are the Single Shot MultiBox Detector (SSD) and You Only Look Once (YOLO) series algorithms, which are extensively used in the recognition and detection of fruits in natural environments due to their faster detection speed and higher scalability [19–21]. Gu et al. [22] proposed a YOLO-DCA model that improves YOLOv7-tiny by using deep separable convolution (DWConv), coordinate attention (CA), and dynamic detection heads, achieving remarkable results in citrus detection. Li et al. [23] designed a lemon target recognition model, Lemon-YOLO, based on the YOLOv3 model, optimizing the model's structure by introducing the SE-ResGNet34 network and the SE-ResNet module while reducing parameters and improving detection speed. Ou et al. [24], based on the YOLOv7 model, added the ShuffleOne module, a slim-neck network, and SimSPPF module operations, effectively detecting passion fruit. Xu et al. [25] significantly improved the recognition accuracy of citrus by incorporating lightweight networks and attention mechanisms into YOLOv4 and adjusting the loss function.

The research mentioned above has promoted the development of fruit recognition and detection technology, but most experimental studies do not consider the actual natural environment. In natural settings, citrus trees have abundant branches and leaves, and the fruits are closely attached. Additionally, branches and leaves often cover each other, complicating detection. Furthermore, the cultivation method of wide-row dense planting with a central trunk results in varying light conditions for citrus fruits. These interfering factors affect the accurate detection of citrus fruits. To address this issue, this study proposes a citrus fruit detection model, YOLO-PBGM, which facilitates the deployment of algorithm models on edge mobile terminals.

This study improves the proposed YOLO-PBGM citrus detection network as follows:

- (1) To address the issues of large model size, high computational complexity, and difficulties in deployment on mobile devices, some CBS modules are replaced with PConv modules to construct the PC-ELAN module. This modification reduces the demand for computational resources and improves inference efficiency on mobile devices.

- (2) The BiFormer attention mechanism is embedded to achieve more flexible calculation allocation and feature perception and to enhance the sensitivity of perception to key data of citrus fruits.
- (3) The GS-ELAN module is constructed using GSConv to provide richer target information and enhance the network's nonlinear capabilities.
- (4) The MDPIoU loss function is utilized to address the problem of distorted detection boxes caused by large sample differences, while also speeding up model convergence.

2. Materials and Methods

2.1. Collecting Datasets

To obtain a dataset that more accurately reflects the state of citrus in a natural orchard environment, data were collected in citrus orchards in Chengdu and Nanchong, Sichuan Province. The orchards utilize a main-stem, wide-row, high-density planting model. The citrus trees exhibit short lateral branches, and the fruit grows mainly on the outside, leaving relatively spacious lateral and bottom spaces on the trunk. This planting structure creates ideal conditions for image acquisition from the side or bottom of the fruit tree. The shooting equipment used for image acquisition was an iPhone 11 (Apple Inc., Cupertino, CA, USA), with an image resolution of 4032×3024 pixels in JPG format. During the shooting process, the handheld device captured images from different distances (30–120 cm) and multiple angles, with the shooting time period ranging from 9:00 to 17:00. Due to the complexity of natural environmental lighting and the high occlusion of citrus fruits, various shooting methods, such as backlighting, were adopted. The collected data samples include images of citrus fruits under different lighting and distance conditions, incorporating occlusion by branches and leaves, overlapping fruits, single fruits, and multiple fruits. Figure 1 shows samples of citrus fruits under different conditions in a complex environment. During the processing of the dataset, images with false shots, high similarity, and high blur were removed, resulting in a total of 407 high-quality original citrus images.

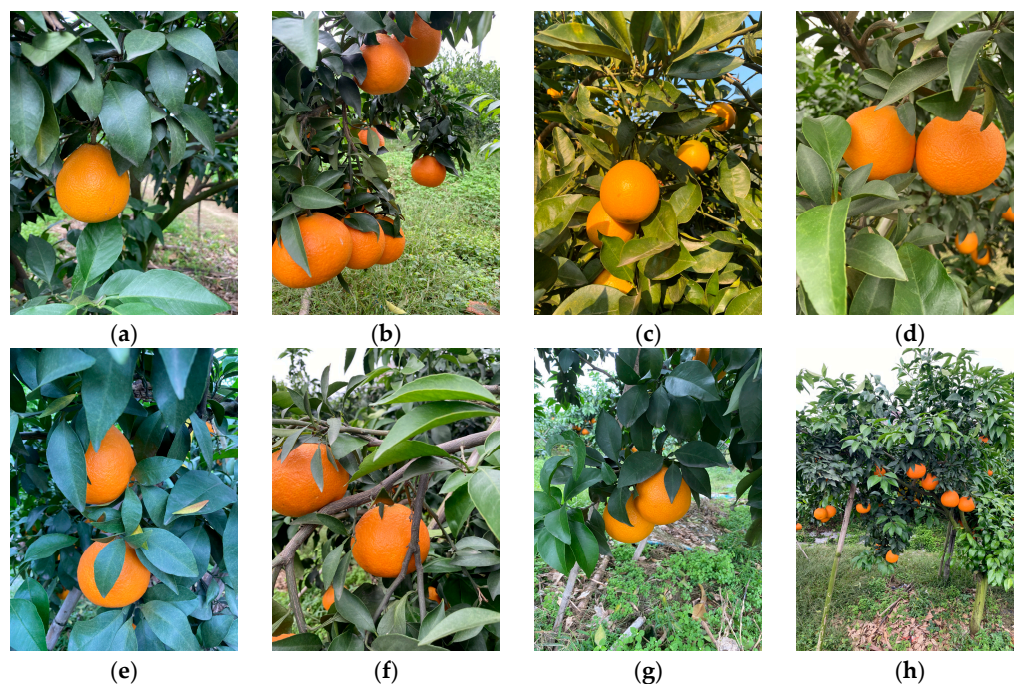


Figure 1. Images of citrus fruit: (a) single fruit; (b) multiple fruits; (c) sunny; (d) overcast; (e) leaves occlusion; (f) stem occlusion; (g) overlap; (h) long distance.

2.2. Dataset Production

Because the data acquisition process is subject to interference, such as light intensity, image clarity, and noise, the collected citrus dataset is augmented so that the model can fully learn the network to fully obtain information about citrus fruits in complex natural scenes, thereby improving its generalization ability in complex environments. This study used a variety of data enhancement methods on the citrus dataset, including brightness adjustment, Gaussian blur, Gaussian noise, mirroring, rotation, and translation, to expand the dataset to 3256 images. Figure 2 displays the images after processing with different data enhancement methods. Additionally, the citrus in the images was manually annotated using Labellmg software (version 1.8.6). The category attribute of the bounding boxes was set to “citrus”, and the annotations were saved to generate a label file in YOLO format.

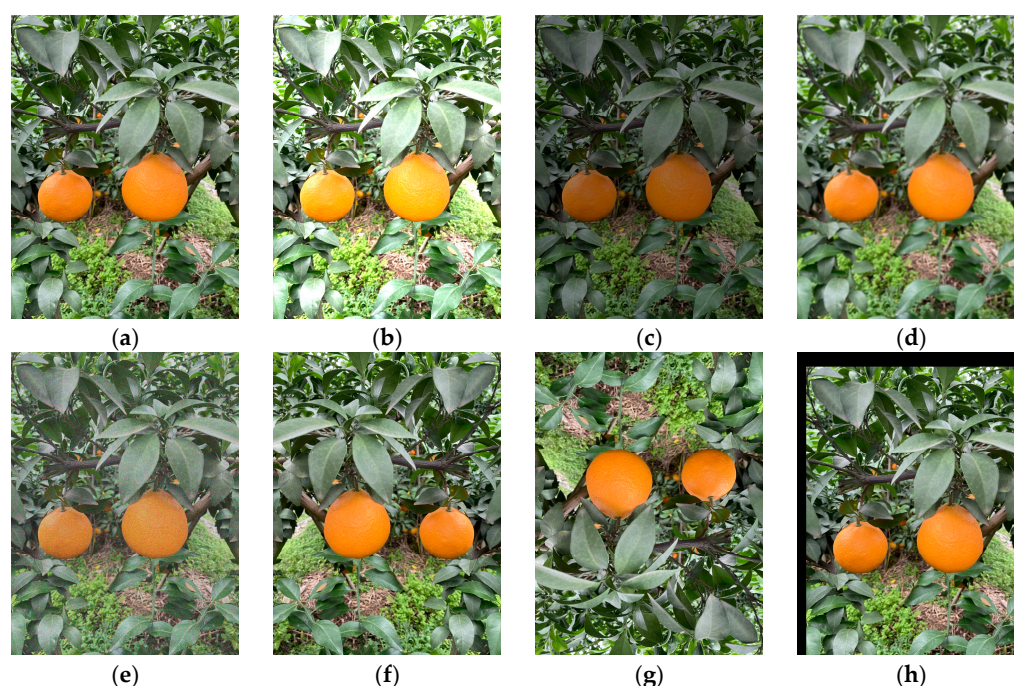


Figure 2. Image enhancement of citrus fruit. (a) Original image; (b) darken; (c) brighten; (d) Gaussian blurring; (e) Gaussian noise; (f) mirroring; (g) rotation; (h) translation.

2.3. YOLOv7 Model

YOLOv7 is a target detection algorithm proposed by the official YOLOv4 team in 2022. Its real-time performance and accuracy surpass those of most similar models currently available [26]. As shown in Figure 3, the YOLOv7 network framework significantly improves model performance by reparametrizing the model’s structure, introducing a strategy for distinguishing between positive and negative samples, and incorporating an auxiliary training mechanism. At the input end, YOLOv7 employs adaptive image scaling and hybrid data enhancement techniques to dynamically adjust the resolution of images to meet the network’s requirements. In the backbone network, ELAN enhances the network’s representation of complex input data by guiding the computational modules in different convolutional groups to learn diverse features. The MPConv layer further improves feature extraction by creating upper and lower branches and fusing features. In the head network, the Spatial Pyramid Pooling, Cross Stage Partial Channel (SPPCSPC) module effectively reduces distortion during image processing and addresses the issue of redundant feature extraction. Additionally, the Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) structures are introduced, which transmit features through an up–down interleaving combination, effectively integrating features at all levels. In the

output layer, the REP structure adjusts the different image channels of the target, and target prediction is performed using a 1×1 convolution. Finally, the model generates feature maps of various sizes to identify objects of different dimensions.

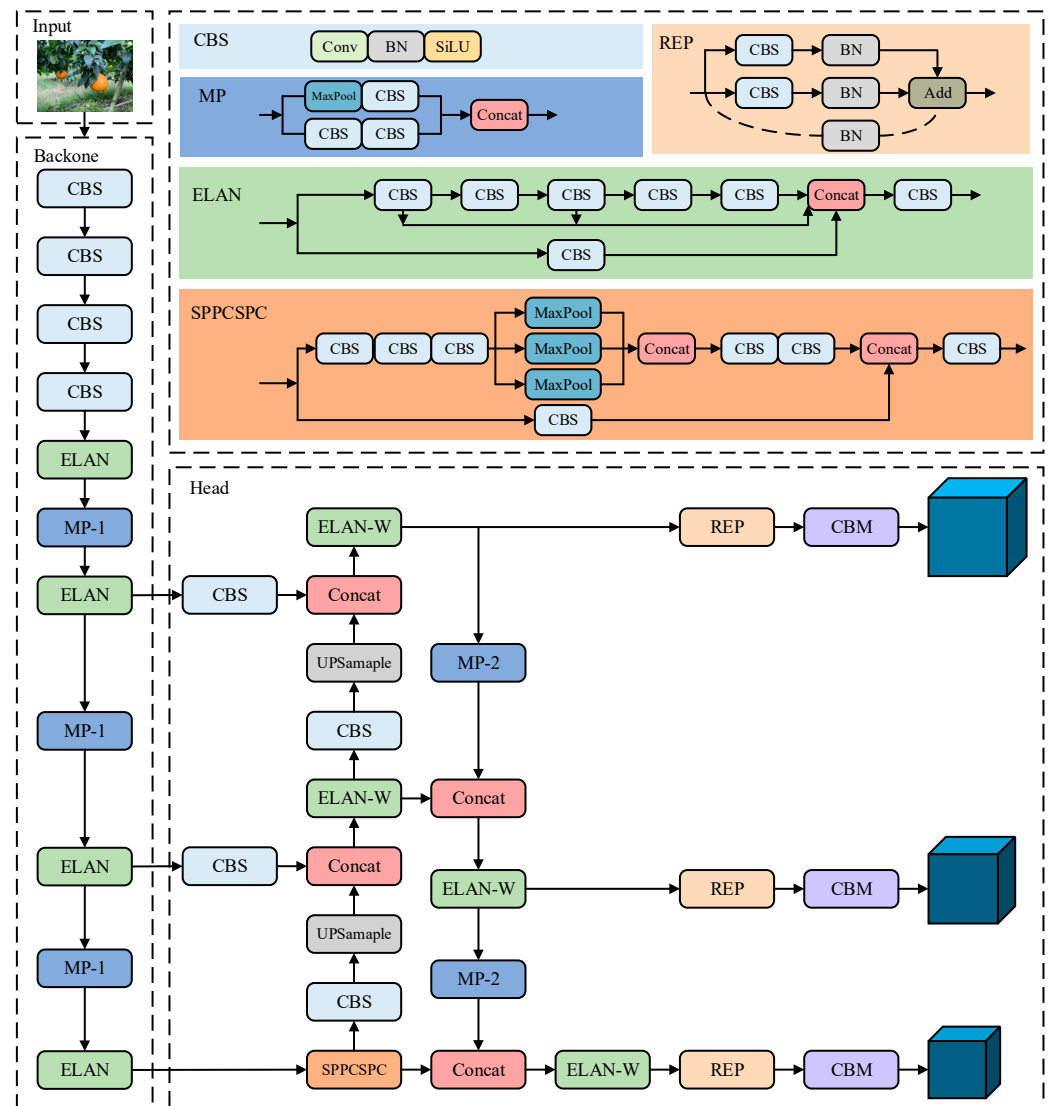


Figure 3. YOLOv7 structure framework.

2.4. YOLO-PBGM Algorithm

The YOLOv7 algorithm has been systematically optimized, resulting in improved performance [27]. However, it remains challenging to effectively detect occluded, dense, or distant citrus targets in complex orchard environments and under uneven lighting conditions. Additionally, the model is relatively large, possesses insufficient feature extraction capability, and contains redundant parameters generated during the training process, which do not enhance the overall efficiency of the model. Therefore, this study selects YOLOv7 for model design and improvement, proposing a more efficient citrus target detection algorithm named YOLO-PBGM. First, Partial Convolution (PConv) is incorporated into ELAN, and the backbone network is redesigned to achieve floating-point operations and streamline computation. Simultaneously, embedded BiFormer attention is utilized for more flexible computational distribution and feature perception. Second, a more lightweight and efficient multi-scale convolution module, GSConv, replaces some of the CBS convolutions in ELAN, thereby constructing the GS-ELAN module, which further improves model performance and reduces computational complexity. Finally, the MPDIoU loss function is

been redesigned, as shown in Figure 5. Additionally, the PConv in FasterNet [29] has been introduced to substitute the CBS in ELAN to construct the PC-ELAN module.

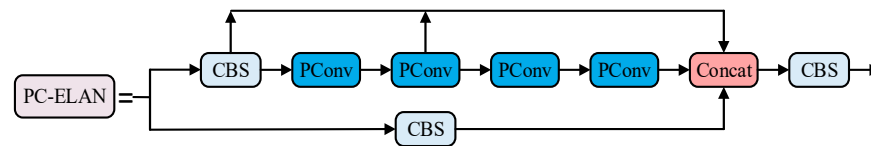


Figure 5. PC-ELAN structure diagram.

PConv can reduce the network's demand for computing resources by decreasing redundant calculations, which effectively lowers the computing load [30]. The working principle is illustrated in Figure 6. Unlike regular convolutions and deep convolutions, PConv utilizes filters to convolve only a small number of input channels, leaving the remaining input channels unaffected. Therefore, while maintaining the same feature extraction capability, PConv achieves lower floating-point operation counts and higher computational speed compared to standard convolutions and deep convolutions.

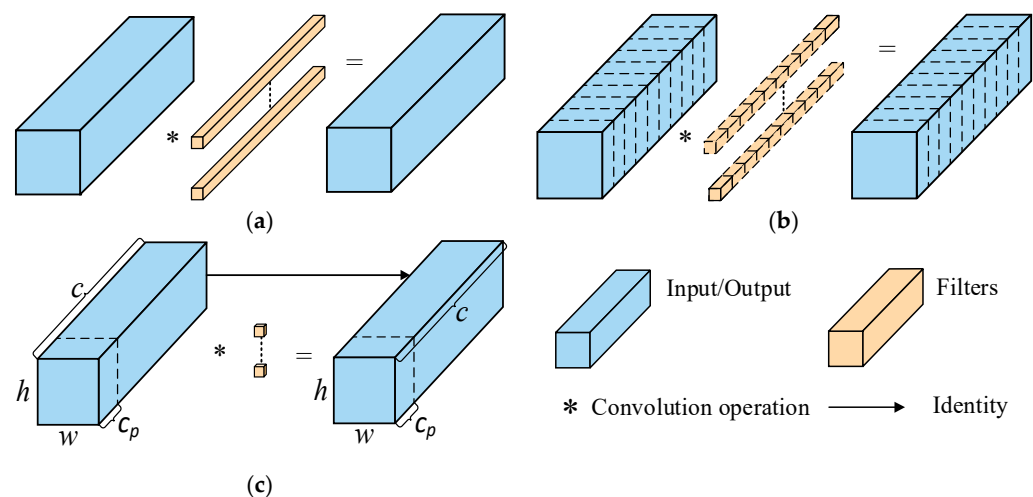


Figure 6. PConv principle of operation. (a) Conventional convolution; (b) deep convolution; (c) Partial Convolution (PConv).

The FLOPs of Pconv are $h \times w \times k^2 \times c_p^2$. By setting the mapping ratio to $r = c_p/c$, the FLOPs can be adjusted. In this study, the mapping ratio is set to 4, and the FLOPs of PConv are only 1/16 of that of the regular convolution.

PConv only participates in spatial feature extraction for certain channels, and there is no loss of information for the subsequent feature channels. This method can effectively prevent operators from over-accessing memory, thereby optimizing the device's processing potential and achieving lower data processing latency.

2.4.2. BiFormer Attention Mechanism

In an orchard environment, citrus fruits are densely distributed and subject to various disturbances, such as mutual occlusion between fruits, which increases the complexity of the citrus detection task. To optimize model performance and enable more effective focus on the key details of the target features, the BiFormer [31] attention mechanism is introduced into the model. BiFormer relies on query adaptation technology, which calculates the correlation between inputs in the feature sequence, ensuring that the model processes global contextual information correctly in complex backgrounds while also enhancing sensitivity to input content.

The BiFormer consists of Bidirectional Relationship Attention (BRA) as its main core module. It captures feature relationships at different scales within the image through its bidirectional modeling capabilities and adaptively adjusts the weights assigned to the features. This approach effectively constructs efficient and powerful data-driven models, making it better equipped to handle complex and large amounts of data. Its structure is depicted in Figure 7a.

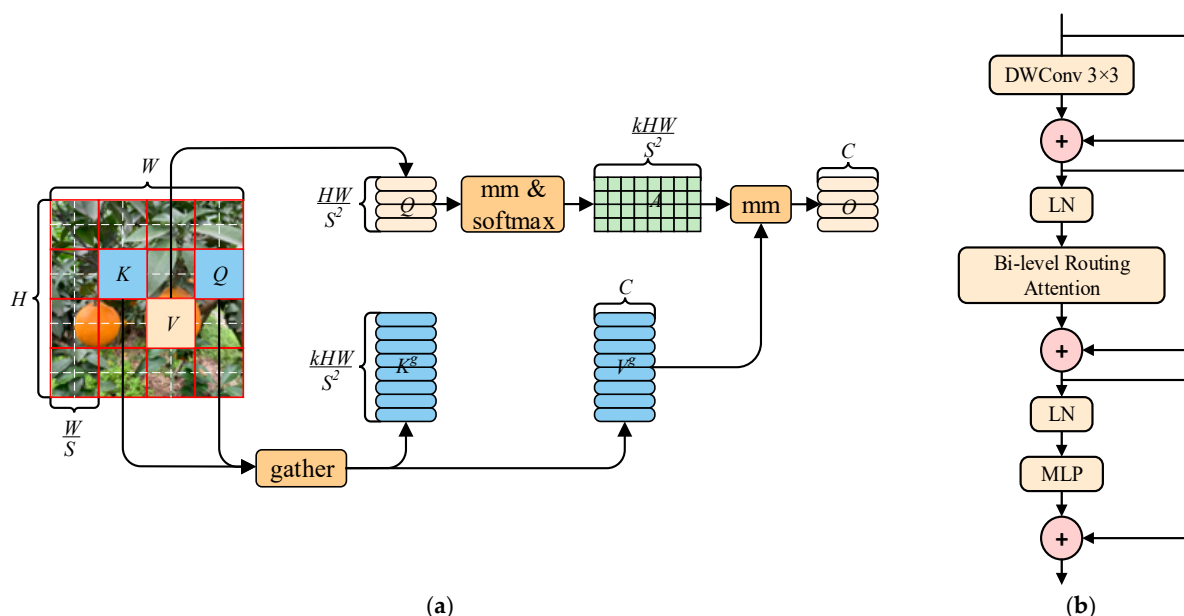


Figure 7. BiFormer model structure. (a) BiFormer; (b) BiFormer Block.

The BiFormer first constructs a region-level directed graph using the target image X input by the network model, and it linearly transforms X to obtain Q , K , and V , as shown in Equation (1).

$$Q = X^r W^q, K = X^r W^k, V = X^r W^v \quad (1)$$

Secondly, we constructed a directed graph to implement the routing between regions. We calculated the average values for Q and K and then performed matrix multiplication to obtain the affinity adjacency matrix between regions, as in Equation (2).

$$A^r = Q^r (K^r)^T \quad (2)$$

Then, only the first k indices of each region were used to cut the target feature map, as in Equation (3), where the i th row of I^r contains the k most relevant indices for the i th region.

$$I^r = \text{topkIndex}(A^r) \quad (3)$$

Finally, the key-value tensors K^S and V^S are collected first, as shown in Equations (4) and (5), and then the attention weights are applied to the aggregated K - V pairs, as indicated in Equation (6).

$$K^S = \text{gather}(K, I^r) \quad (4)$$

$$V^S = \text{gather}(V, I^r) \quad (5)$$

$$O = \text{Attention}(Q, K^S, V^S) + \text{LCE}(V) \quad (6)$$

With the BiFormer Block, as shown in Figure 7b, the use of DWConv can reduce the resources required for memory reading and writing, shortening the actual network computing time. In addition, the use of LN can accelerate model training and improve

model generalization performance. Multilayer perceptrons (MLPs) are used to adjust and optimize weights in greater depth, enabling the model to capture the target's complex feature relationships.

2.4.3. GSConv Convolution

In the input module, citrus fruits are densely distributed throughout the image and exhibit low resolution. However, the stacked 3×3 convolutional layers in the CBS structure have certain limitations in feature extraction. The fixed receptive field restricts each convolutional layer's ability to capture local details of small targets, thereby reducing the model's perceptual ability [32]. Furthermore, stacking too many CBS modules increases the computational burden of the network and hinders deployment on small mobile devices. This study aims to maintain a balance between accuracy and model size, achieving better detection accuracy using limited computing resources. Currently, lightweight network design primarily involves reducing model parameters and floating-point operations by utilizing deep separable convolution (DSCConv) operations [33]. Standard convolution is a channel-intensive operation, while the DSCConv operation applies convolution across three channels separately to minimize redundancy in feature information. However, this approach also separates the channel information of the input image, which degrades the network's object feature perception ability. Therefore, the GSConv [34] module is introduced, replacing standard convolution with GSConv. GSConv incorporates lightweight concepts from GhostNet and ShuffleNetv2 to capture both local and global structural features by aggregating multi-layer neighbor information, thereby effectively capturing global information. Its structure is illustrated in Figure 8.

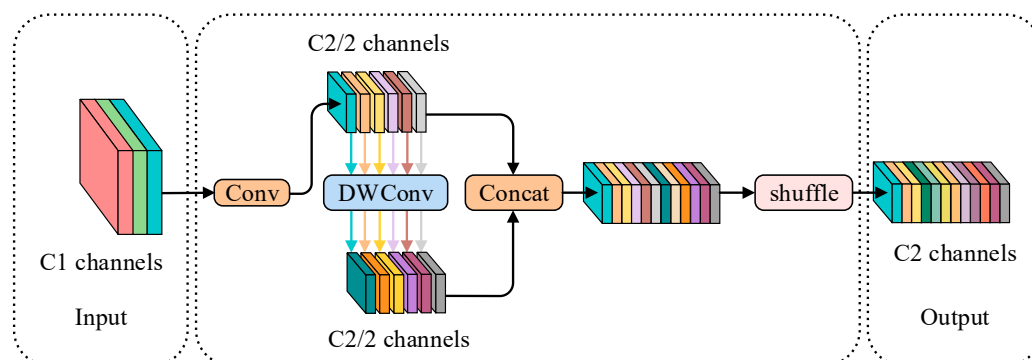


Figure 8. Schematic diagram of GSConv module structure.

GSConv passes the feature map with the input channel number $C1$ through the convolution module first to form two sets of sub-feature maps. The two subsets are then directed to different processing paths, each receiving specific convolution operations in independent sub-channels. Half of the feature maps are processed using depth-separable convolutions, while the other half utilizes standard convolutions. The convolution results from the sub-channels are then integrated and rearranged to combine the two-channel feature maps. Finally, the shuffle module transmits its target information through each channel to improve the flow of information between features.

If the GSConv module replaces all of the CBS modules in ELAN, the model becomes excessively large and complex, increasing the resistance of the data flow and prolonging inference time, which leads to a decrease in the frame rate per second. Accordingly, the ELAN in the YOLOv7 neck network was redesigned, as shown in Figure 9, and the GS-ELAN module was constructed by using GSConv instead of CBS in ELAN.

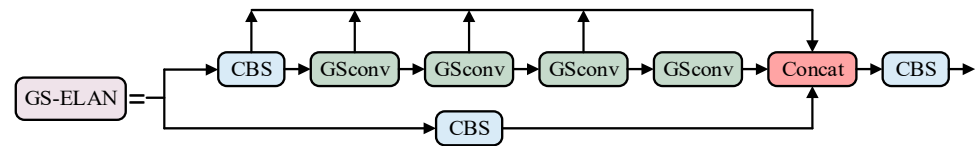


Figure 9. Schematic structure of the GS-ELAN multi-scale feature processing module.

2.4.4. MPDIoU Loss Function

During the automated harvesting of citrus fruits, the heavy obscuration of the fruits by branches and leaves, along with the overlapping nature of the fruits, pose significant challenges for harvesting robots in identifying the fruits. YOLOv7 regression loss employs CIOU loss [35]. Although Complete Intersection over Union (CIOU) can effectively describe the regression state of the bounding box, the relative scale penalty term v is less effective when dealing with occluded and overlapping targets. Additionally, the distance between the two target boxes cannot be accurately determined, leading to a higher likelihood of missed detections during non-maximum suppression, which subsequently reduces the accuracy of bounding box regression.

To more effectively address the challenges of bounding box regression for citrus fruits in occluded and overlapping situations, this study introduces MPDIoU. The schematic diagram is shown in Figure 10. MPDIoU is based on a similarity measure of the minimum point distance of the bounding box, which can better handle the impact of overlapping objects on the bounding box. Furthermore, because the detection task for citrus targets requires consideration of factors like shape and orientation, MPDIoU [36] introduces parameters like diagonal length. This significantly improves the geometric alignment ability of the bounding box, making boundary fitting more accurate while reducing dependence on intersection and comparison. The calculation is as follows.

$$d_1^2 = (x_1^B - x_1^A)^2 + (y_1^B - y_1^A)^2 \quad (7)$$

$$d_2^2 = (x_2^B - x_2^A)^2 + (y_2^B - y_2^A)^2 \quad (8)$$

$$MPDIoU = \frac{A \cap B}{A \cup B} - \frac{d_1^2}{w^2 + h^2} - \frac{d_2^2}{w^2 + h^2} \quad (9)$$

where w and h are the width and height of the input image, A and B are the predicted and real frames, respectively, (x_1^A, y_1^A) , (x_1^B, y_1^B) are the upper-left coordinates of bounding boxes A and B , respectively, and (x_2^A, y_2^A) , (x_2^B, y_2^B) are the lower-right coordinates of bounding boxes A and B , respectively.

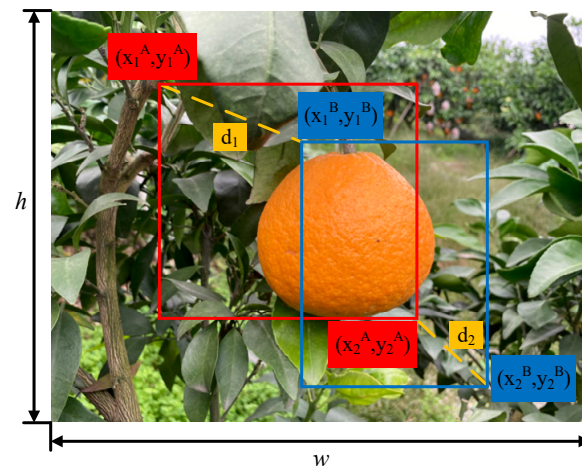


Figure 10. MPDIoU schematic diagram.

3. Results

3.1. Experimental Platform

A server was used for model training and optimization during the experiments, with Windows 10 as the operating system, AMD Ryzen 97950X3D 16-Core Processor 4.20GHz (Advanced Micro Devices, Inc., Santa Clara, CA, USA) as the processor, NVIDIA GeForce RTX 4090D (Nvidia, Santa Clara, CA, USA) as the graphics card, and 64GB of RAM running on the computer. PyTorch1.11.0 was used Python 3.9, CUDA 11.8, and cuDNN 8.9.5 as the framework for developing the deep learning model. The model weights file was saved once after every 25 rounds of training. Table 1 shows the training strategy.

Table 1. Hyperparameter configurations for citrus target detection model training.

Parameter Name	Parameter Value
Image size	640 × 640
Batch size	16
Multi-threaded	16
Momentum	0.937
Initial learning rate	0.01
Optimizer	SGD
Epochs	300

3.2. Evaluation Index

The proposed model was evaluated using the following metrics: precision (P), recall (R), mean average precision (mAP), frames per second (FPS), memory footprint (Params), and GFLOPS [37]. The calculation formulas are as follows.

$$P = \frac{TP}{TP + FP} \quad (10)$$

$$R = \frac{TP}{TP + FN} \quad (11)$$

where TP indicates the samples that were correctly judged as citrus by the target detection model, FP indicates the samples that were judged as citrus by the target detection model, and FN is the citrus sample that was not recognized by the target detection model.

$$mAP = \frac{\sum_{i=1}^{N_c} AP_i}{N_c} \quad (12)$$

where N_c represents the number of all objects in the target dataset, $mAP@0.5$ is the average detection accuracy of citrus targets when the IoU weight is 0.5, and AP_i is the average accuracy of the i th category, which can comprehensively consider the impact of both accuracy and recall.

3.3. Attention Mechanism Comparison Experiment

Five comparative experiments were designed to introduce the Convolutional Block Attention Module (CBAM) [38], Efficient Channel Attention (ECA) [39], Squeeze-and-Excitation (SE) [40], the Similarity-based Attention Module (SimAm) [41], and BiFormer attention modules to modify the YOLOv7 network and compare the detection performance on the same citrus dataset using the same server platform. Table 2 presents the results of model detection. The analysis indicates that when different attention mechanisms are added to YOLOv7, the model's parameters and GFLOPs remain largely unchanged. Among these mechanisms, the model's metrics were most improved with the inclusion of BiFormer

in the network structure, resulting in a 3.2% increase in mAP and the highest frame rate (FPS). The BiFormer effectively models long-range relationships during image processing, enhancing the model's capacity to capture contextual dependencies over longer distances. By optimizing the fusion and processing of citrus fruit features, it facilitates more accurate detection of citrus fruits.

Table 2. Detection results of models that introduce different attention mechanisms.

Model	P (%)	R (%)	mAP@0.5 (%)	Params (M)	GFLOPs (G)	FPS (f·s ^{−1})
Baseline	98.2	93.0	92.7	37.20	105.1	60.24
CBAM	98.4	94.0	93.5	37.36	105.4	64.52
ECA	97.1	94.0	93.2	37.59	106.4	60.61
SE	96.8	97.0	95.3	37.39	105.3	67.11
SimAm	96.4	97.0	95.5	37.31	105.1	61.35
BiFormer	98.5	97.0	95.9	37.28	105.1	64.96

3.4. Comparative Experiment of Loss Functions

To verify the effectiveness of the MPDIoU loss function in citrus object detection, the Distance-IoU (DIoU), Expected Intersection over Union (EIoU) [42], Soft Intersection over Union (SIoU) [43], and MPDIoU loss functions were introduced into the YOLOv7 network. Comparative experiments were conducted using the same experimental platform and dataset. Table 3 and Figure 11 present the detection performance and loss values for each loss function. Table 3 demonstrates that on the citrus dataset, MPDIoU increases the mAP of the original CIoU loss function by 1.8% and the FPS by 3.86 frames. Additionally, MPDIoU outperforms other loss functions across all performance metrics.

Table 3. Comparison of loss function results.

IoU Loss	P (%)	R (%)	mAP@0.5 (%)	Params (M)	GFLOPs (G)	FPS (f·s ^{−1})
CIoU	98.2	93.0	92.7	37.20	105.1	60.24
DIoU	98.1	95.0	93.7	37.21	105.1	58.82
GIoU	98.6	94.0	92.8	37.20	105.2	54.05
SIoU	98.2	94.0	92.9	37.20	105.1	61.73
MPDIoU	98.5	97.0	94.5	37.20	105.1	64.10

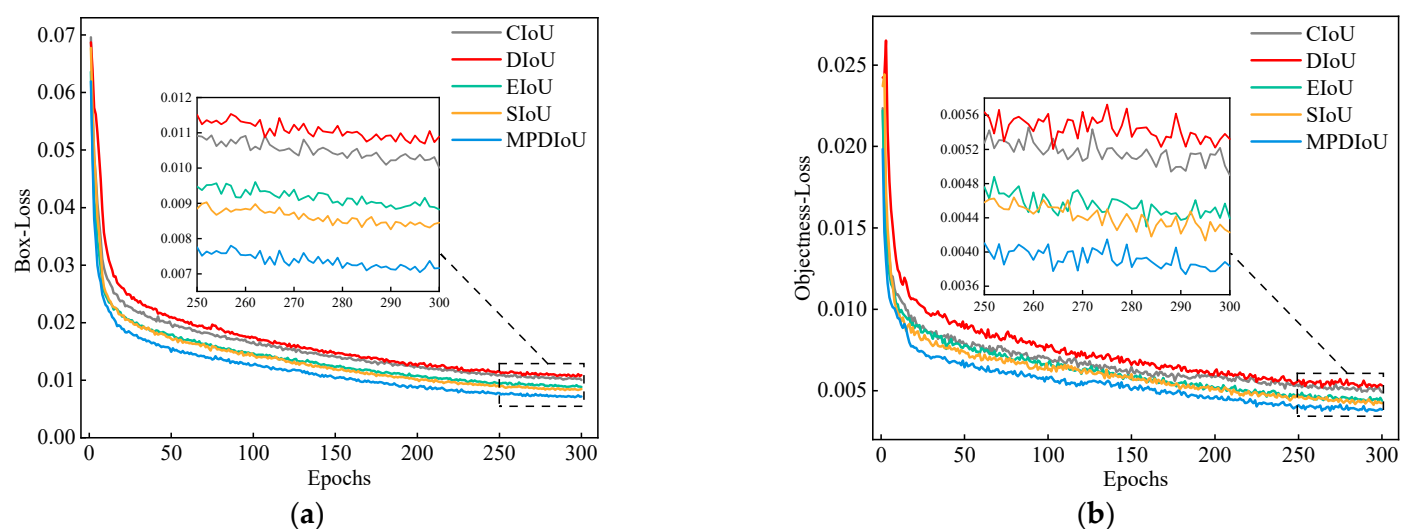


Figure 11. Variation curves of different loss functions. (a) Location loss; (b) confidence loss.

Figure 11 illustrates that the loss values of each loss function decrease as the training process progresses, entering a stable state in the later stages. When using CIoU and DIoU,

the model's loss function converges the slowest, with the loss value remaining at a high level throughout the training process. EIoU and SIoU exhibit similar performance, showing little difference in convergence speed and final loss value. The MPDIoU loss function proposed in this study demonstrates the best performance in the experiments. Experimental data indicate that when utilizing MPDIoU, the gradient of the model decreases significantly faster during the early stages of training compared to other loss functions, beginning to converge stably after approximately 250 rounds. After convergence, the loss value of MPDIoU remains significantly lower than that of other models, demonstrating superior optimization capabilities and accuracy. Therefore, MPDIoU not only accelerates model convergence but also achieves superior final performance, showcasing strong practical application potential.

3.5. Ablation Experiment

In this experiment, YOLOv7 serves as the baseline model, and the added modules are analyzed in terms of individual contributions and combinations of multiple modules. The results of the ablation experiments are presented in Table 4. Table 4 demonstrates that adding the PConv module to YOLOv7 improves the model's mAP by 1%, while reducing Params and GFLOPs by 12.1% and 19.3%, respectively. The introduction of the BiFormer module results in increases of 3.2% in mAP and 7.8% in FPS. When the GSConv module is utilized, the model's mAP increases by 1.3%, with Params and GFLOPs decreasing by 6.1% and 4.3%, respectively. By optimizing regression efficiency for citrus targets, MPDIoU enhances the model's mAP by 1.8%. Additionally, the combination of the BiFormer and PConv modules not only improves detection accuracy but also achieves a lighter model. When the PConv, BiFormer, and GSConv modules are employed together, the model's mAP and FPS increase by 2.6% and 24.8%, respectively. Integrating all four improvement modules results in optimal comprehensive detection capability, with the YOLO-PBGM network achieving an mAP of 96.2% and an FPS of 74.46. In summary, the experiment further verifies the enhancement of the model through the four improved modules across multiple indicators. Meanwhile, it reduces Params and conserves computing resources, thereby decreasing time consumption and providing a feasible solution for deploying the model as a target detection device on picking robots.

Table 4. Results of ablation experiments.

Baseline	PConv	BiFormer	GSConv	MPDIoU	mAP@0.5 (%)	Params (M)	GFLOPs (G)	FPS (f·s ⁻¹)
YOLOv7	×	×	×	×	92.7	37.20	105.1	60.24
	✓	×	×	×	93.7	32.69	84.8	70.92
	×	✓	×	×	95.9	37.28	105.1	64.96
	×	×	✓	×	94.0	34.92	100.5	66.23
	×	×	×	✓	94.5	37.20	105.1	64.10
	✓	✓	×	×	95.1	33.74	84.8	74.46
	✓	✓	✓	×	95.3	31.47	80.2	75.19
	✓	✓	✓	✓	96.2	31.47	80.2	77.51

Note: “×” denotes no module added; “✓” indicates use of the model.

3.6. Improved YOLO-PBGM Comparison Experiment

To verify the superiority of the YOLO-PBGM citrus fruit detection network for target recognition of citrus fruits in an orchard environment, an experiment was conducted to compare the YOLOv7 and YOLO-PBGM detection networks on a citrus dataset while maintaining uniform environmental and training strategies. Table 5 shows that compared to YOLOv7, the YOLO-PBGM detection network achieved an accuracy rate that was 0.3% higher for the recognition of citrus fruits, a recall rate that was 4% higher, and a mean

average precision (mAP) that was 3.5% higher, indicating more accurate detection. In addition, by optimizing the model's architecture, the parameters (Params) and GFLOPs of YOLO-PBGM were reduced by 15.4% and 23.7%, respectively, while the frames per second (FPS) increased by 17.27 f/s. The YOLO-PBGM network reduces the demand for computing resources and storage space, speeds up detection, and ensures the efficiency of citrus target detection in complex environments.

Table 5. Detection performance before and after model improvement.

Model	P (%)	R (%)	mAP@0.5 (%)	Params (M)	GFLOPs (G)	FPS (f.s ⁻¹)
YOLOv7	98.2	93.0	92.7	37.20	105.1	60.24
YOLO-PBGM	98.5	97.0	96.2	31.47	80.2	77.51

Figure 12 visually demonstrates the changes in the loss function and accuracy before and after model improvement. The YOLO-PBGM network significantly accelerates the convergence of the boundary regression process, and both its position loss and confidence loss exhibit lower loss values during convergence. An analysis of the recall and mAP curves indicates that the improved model achieves higher accuracy in the citrus object detection task. Therefore, the proposed YOLO-PBGM model demonstrates stronger stability and excellent detection performance.

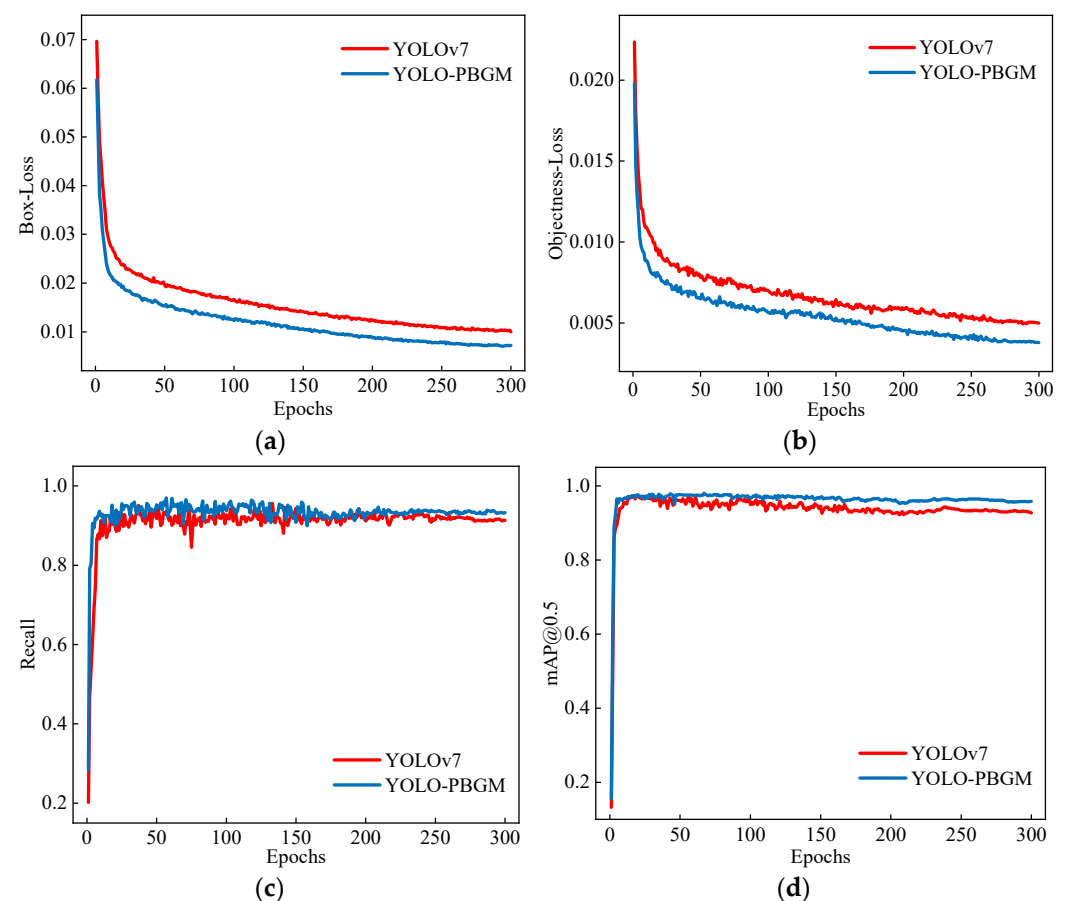


Figure 12. Model loss function and accuracy change curve. (a) Location loss function; (b) confidence loss function; (c) recall; (d) mAP@0.5.

Figure 13 illustrates the recognition performance of YOLOv7 and YOLO-PBGM for citrus in various scenarios. The white boxes in the figure indicate citrus targets that were missed by the models. Both models demonstrate good detection performance under

backlight and low-light conditions. However, the YOLOv7 algorithm struggles with dense targets and small distant targets, resulting in the omission of some partially occluded citrus targets. This is primarily because the YOLOv7 model does not adequately focus on the characteristics of citrus, making it difficult to effectively distinguish the fruit from the background in complex environments. In contrast, the YOLO-PBGM model effectively reduces the miss rate and exhibits higher recognition confidence across different scenarios. The superior detection performance of YOLO-PBGM is evidenced by its successful detection of citrus targets in an orchard environment.

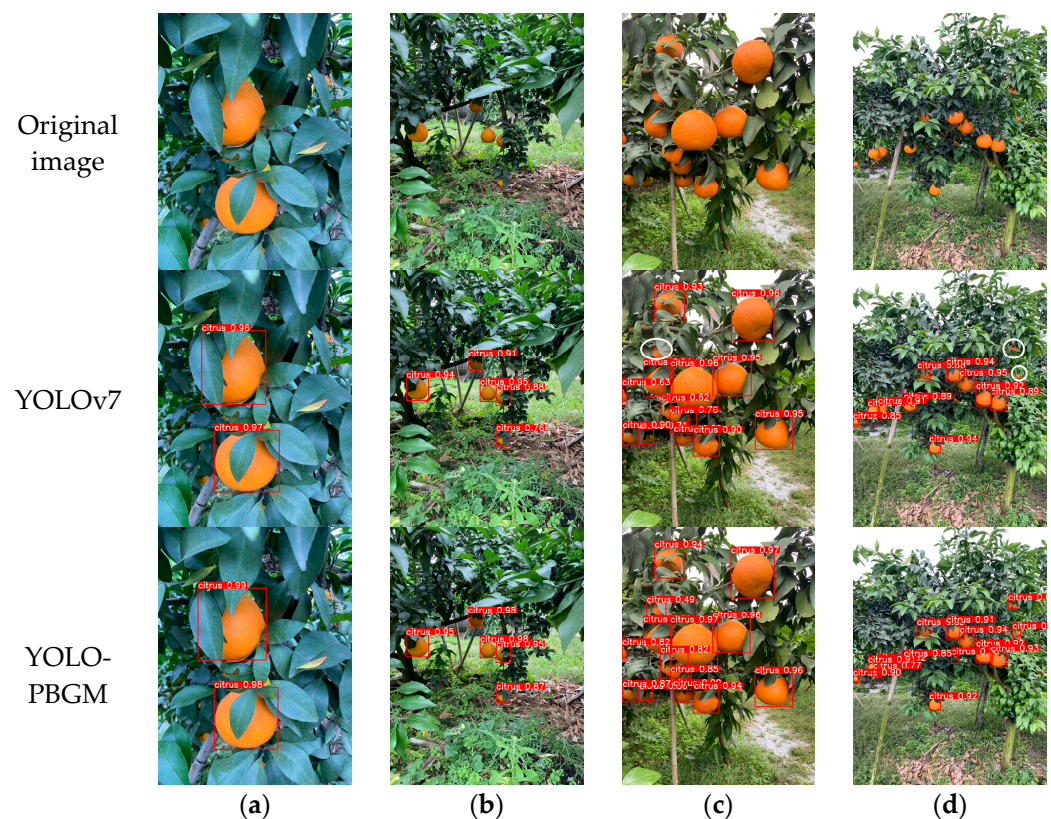


Figure 13. Model detection effect. (a) Front light; (b) backlight; (c) intensive; (d) long distance.

3.7. Model Comparison

3.7.1. Detection Performance Analysis with Benchmark Model

To objectively demonstrate the advantages of the YOLO-PBGM network model, comparisons were conducted with the mainstream network models Faster R-CNN, YOLOv5 [44], YOLOv6 [45], YOLOv7, and YOLOv8 [46]. To ensure fairness and reproducibility in the model comparison, all comparisons in this study strictly adhered to the parameter configurations shown in Table 1, and we fixed the number of training iterations to 300. These standardized settings eliminate the impact of differences in training strategies on performance evaluation and ensure that the experimental results objectively reflect the performance differences of the model architectures themselves. The performance comparison results of different mainstream detection models are presented in Table 6. Analyzing Table 6 reveals that in terms of accuracy, the YOLO-PBGM algorithm achieves a mean average precision (mAP) of 96.2%. This represents improvements of 7.1%, 4.4%, 4.7%, 3.5%, and 2.8% compared to Faster R-CNN, YOLOv5, YOLOv6, YOLOv7, and YOLOv8, respectively. In terms of computational complexity, YOLO-PBGM demonstrates a clear advantage. The model's parameters total 31.47 M, which is 77%, 31.8%, 9.8%, 15.4%, and 27.9% lower than those of the five aforementioned models, respectively. Additionally, the number of floating-point operations is 80.2 G, which is significantly lower than those of the

other algorithms. Regarding real-time performance, the YOLO-PBGM algorithm achieves an FPS of 77.51, showing improvements of 56.15 FPS, 23.46 FPS, 23.89 FPS, 17.27 FPS, and 13.93 FPS compared to the five aforementioned models. Considering the indicators of real-time performance, accuracy, and computational complexity, the performance of the YOLO-PBGM algorithm is significantly superior to that of the other algorithms. Furthermore, it reduces the memory size and computational consumption of the neural network model, providing a technical reference for the deployment of algorithm models for citrus detection on edge mobile terminals.

Table 6. Performance comparison of different mainstream detection models.

Model	P (%)	R (%)	mAP@0.5 (%)	Params (M)	GFLOPs (G)	FPS (f·s ^{−1})
Faster R-CNN	87.9	88.6	89.1	136.75	368.3	21.36
YOLOv5	94.7	92.5	91.8	46.14	108.2	54.05
YOLOv6	93.8	91.7	91.5	34.87	85.3	53.62
YOLOv7	98.2	93.0	92.7	37.20	105.1	60.24
YOLOv8	97.6	93.2	93.4	43.63	165.4	63.58
YOLO-PBGM	98.5	97.0	96.2	31.47	80.2	77.51

To compare the recognition effectiveness of different algorithms for citrus, experiments were conducted using a citrus target image dataset. The detection performance of various models is illustrated in Figure 14. The white boxes indicate citrus targets that were missed by the models, while the red boxes highlight instances where the models mistakenly recognized multiple citrus targets as a single target. Figure 14 reveals the missed detection phenomenon present in all five mainstream network models. Specifically, when the fruits overlap, the unclear boundaries between the mandarins often lead to misidentification, causing the models to erroneously classify several mandarins that are stuck together as one target, resulting in inaccurate recognition outcomes. Compared to the other models, YOLO-PBGM demonstrates superior recognition capability in addressing leaf occlusions and overlapping mandarins, yielding better detection results in actual complex environments.

3.7.2. Comparison of Different Citrus Detection Models

To further validate the performance advantages of the YOLO-PBGM model in the citrus target detection task, various current citrus target detection models were selected for comparative experiments in this study. As shown in Table 7, the YOLO-PBGM model significantly outperforms the comparison models in three key metrics, precision, recall, and mean average precision (mAP), demonstrating excellent detection performance. Specifically, regarding detection precision, the YOLO-PBGM model achieves a precision rate of 98.5%, significantly surpassing the YOLO-DCA and YOLOv7-BiGS models. This indicates a strong anti-interference capability in complex orchard environments, effectively reducing misdetection issues caused by background noise, branch and leaf occlusion, and other factors. In terms of recall, the YOLO-PBGM model shows a 5.4% improvement compared to the YOLO-DCA model, demonstrating its excellent recognition ability for citrus fruits under varying growth states, shading levels, and light conditions. The comprehensive performance index mAP is 96.2%, reflecting an average improvement of approximately 7.2% over the comparison models, thereby illustrating the synergistic optimization effect of the model on target localization and classification tasks.

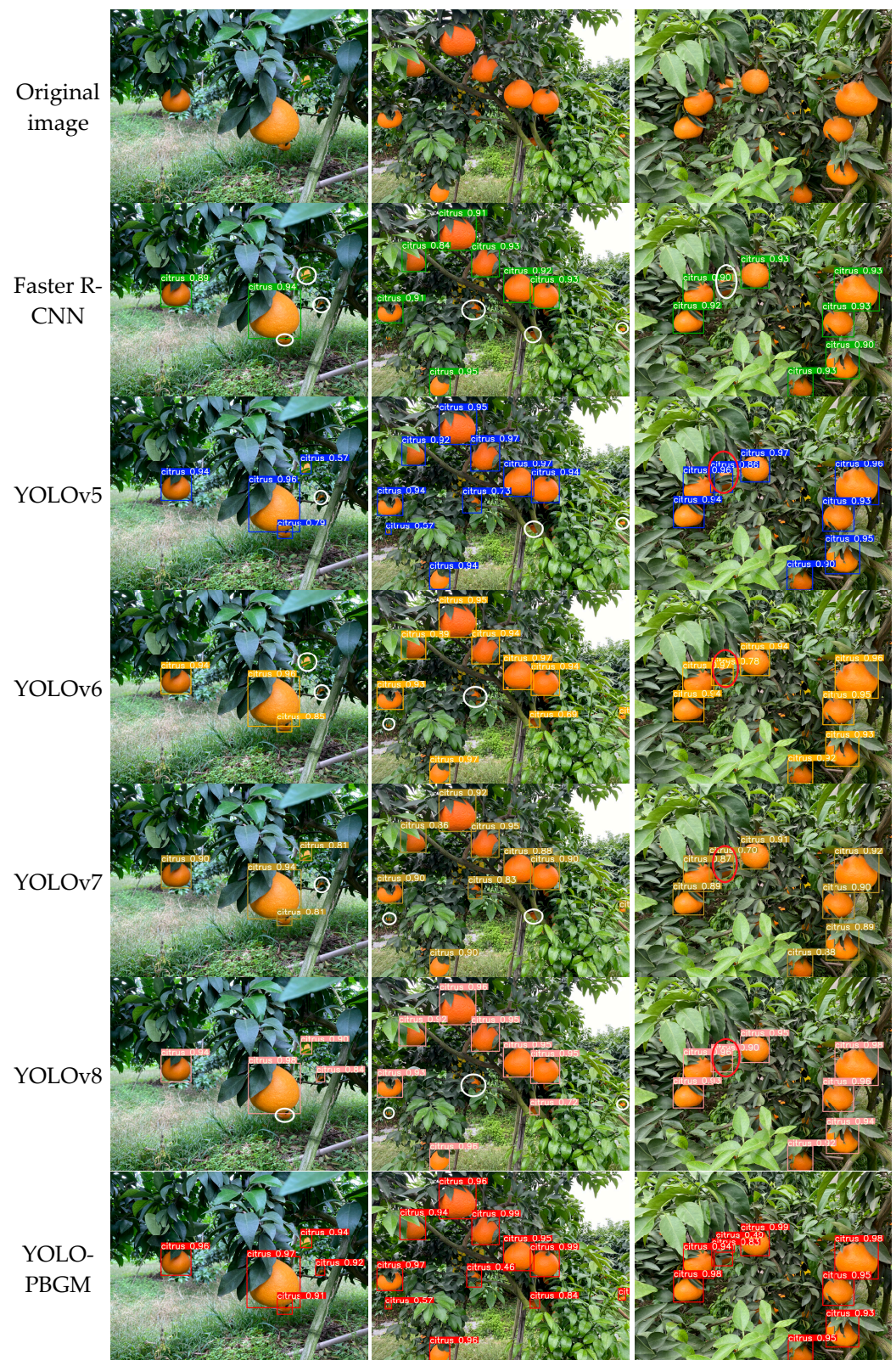


Figure 14. Detection effect of different mainstream models.

Table 7. Performance comparison of different citrus detection models.

Model	P (%)	R (%)	mAP@0.5 (%)
YOLO-BP [47]	86.0	91.0	91.6
YOLO-DCA [22]	94.1	91.6	95.0
AG-YOLO [48]	90.6	73.4	83.2
YOLOv7-BiGS [49]	91.0	87.3	93.7
YOLO-MECD [50]	84.4	73.3	81.6
YOLO-PBGM	98.5	97.0	96.2

4. Discussion

In the natural environment, the mechanized harvesting of citrus fruits faces complex environmental factors and a variety of fruit conditions. To more accurately evaluate the robustness of the algorithm and further assess its effectiveness, a citrus target dataset was tested under different lighting conditions and occlusions.

4.1. Detection of Different Lighting Conditions

To study the robustness of the YOLO-PBGM network under different lighting conditions in an orchard environment, this experiment selected images of citrus fruits taken in sunny and cloudy conditions for testing. Table 8 presents the detection results under these varying lighting conditions. The analysis indicates that under sunny conditions, the YOLO-PBGM model achieved a mean average precision (mAP) of 96.2% and a frames per second (FPS) rate of 76.52 f/s, while, under cloudy conditions, the model yielded an mAP of 95.7% and an FPS of 75.26 f/s. In summary, although the mAP and FPS of the YOLO-PBGM model vary under different lighting conditions, its overall performance remains at a high level. Figure 15 displays detection images of citrus under these different lighting conditions. The results demonstrate that the model exhibits a low missed detection rate for citrus targets, high recognition accuracy, and robust recognition capabilities across varying lighting conditions.

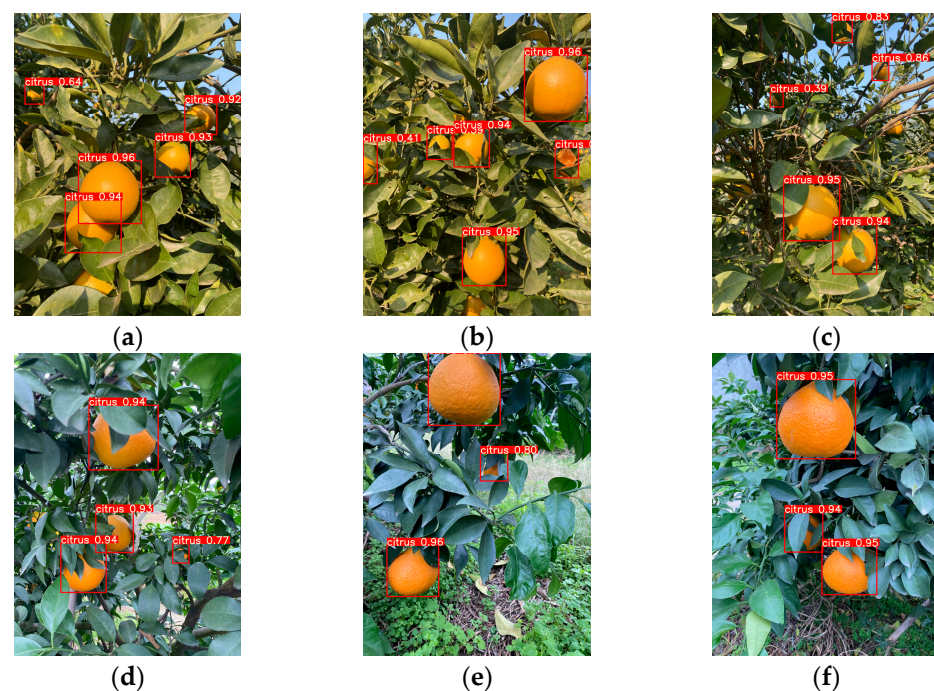
**Figure 15.** Detection results under different lighting conditions. (a–c) Sunny; (d–f) overcast.

Table 8. Detection results under different lighting conditions.

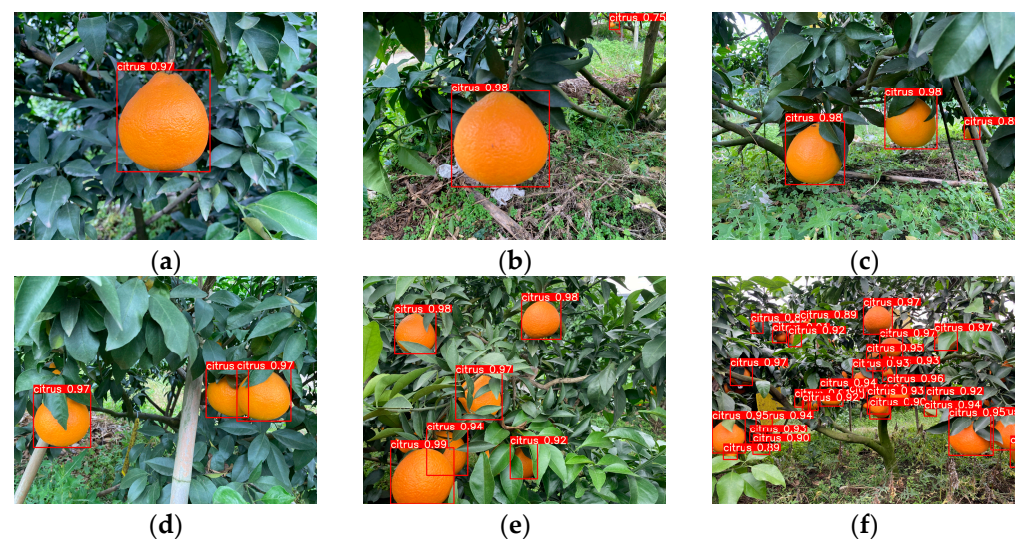
Model	Light	P (%)	R (%)	mAP@0.5 (%)	FPS (f·s ^{−1})
YOLO-PBGM	Sunny	97.6	96.5	96.2	76.52
	Overcast	97.2	95.8	95.7	75.26

4.2. Detection of Different Occlusion Conditions

To study the detection performance of the YOLO-PBGM model in an orchard environment for citrus fruit occlusion, this experiment selected three different levels of fruit occlusion scenarios for evaluation: no occlusion, slight occlusion, and severe occlusion. Table 9 presents the detection results under different occlusion conditions. The experimental results of occluded target detection are shown in Table 9. The analysis reveals that the mean mAPs for the three occlusion conditions are 99.6%, 97.2%, and 92.2%, respectively, while the FPSs are 79.35 f/s, 76.18 f/s, and 73.49 f/s, respectively. Figure 16 illustrates the detection image of occluded citrus by the YOLO-PBGM model, in which all citrus fruits are successfully detected. This demonstrates that the YOLO-PBGM model can effectively learn the key features of citrus targets, addressing the issue of feature information loss caused by occlusion and exhibiting superior occlusion resistance and stability in the detection of occluded citrus.

Table 9. Detection results for different occlusion conditions.

Model	Type	P (%)	R (%)	mAP@0.5 (%)	FPS (f·s ^{−1})
YOLO-PBGM	Unobstructed	98.9	99.1	99.6	79.35
	Slightly obscured	98.2	97.6	97.4	76.18
	Severely obstructed	96.4	95.3	92.2	73.49

**Figure 16.** Detection results under different occlusion conditions. (a,b) Unobstructed; (c,d) slightly obscured; (e,f) severely obstructed.

5. Conclusions

To address the poor adaptability and high computational complexity of the model for detecting citrus fruits in natural scenes, this study improves the YOLOv7 model and proposes the YOLO-PBGM network model, which considers both inference speed and accuracy. The model is enhanced by introducing PConv and GSConv to replace some CBS modules, thereby reducing computational complexity and model parameters. Additionally, the BiFormer attention mechanism is embedded to achieve more flexible calculation

allocation and feature perception, enhancing the network's focus on citrus fruit features. Furthermore, the MPDIoU loss function is added to accelerate model convergence and reduce regression loss. The precision, recall, and mAP of the model for citrus identification in complex environments are 98.5%, 97%, and 96.2%, respectively. Compared to mainstream network models, the YOLO-PBGM model demonstrates the best overall performance on the citrus dataset in an orchard environment, effectively reducing missed detections of citrus targets. In comparison to YOLOv7, the model parameters and floating-point calculations have been reduced by 15.4% and 23.7%, respectively, while the FPS has increased by 17.27 f/s, facilitating deployment on mobile devices. Meanwhile, experiments on instance recognition in complex scenarios with varying lighting and occlusion conditions have demonstrated good detection performance and high robustness, providing strong support for citrus picking recognition in complex environments.

The YOLO-PBGM model exhibits superior performance in citrus detection within complex environments, but there remains potential for improvement. The current dataset comprises images of mature citrus fruits with orange skin, lacking samples of green citrus in the immature stage and yellow citrus in the semi-mature stage. This limitation may reduce the model's performance in identifying citrus fruits of varying colors and maturities. Consequently, subsequent research will introduce images of citrus fruits at different growth stages to enrich the sample dataset, thereby enhancing the model's versatility and adaptability in detecting citrus fruits across different growth stages.

Author Contributions: Conceptualization, Q.L. and F.S.; methodology, Q.L. and F.S.; software, Q.L. and Y.B.; validation, Y.B. and H.W.; formal analysis, X.L. (Xiaoxiao Li); investigation, J.Z.; resources, X.L. (Xin Li); data curation, Y.B.; writing—original draft preparation, Q.L.; writing—review and editing, Q.L., Y.B., and H.W.; funding acquisition, F.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Sichuan Provincial Department of Science and Technology (project number: 2024YFHZ0147) and the Chengdu University National College Student Entrepreneurship Training Program (project number: 202411079004S).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Hayat, F.; Li, J.; Iqbal, S.; Peng, Y.; Hong, L.; Balal, R.M.; Khan, M.N.; Nawaz, M.A.; Khan, U.; Farhan, M.A. A mini review of citrus rootstocks and their role in high-density orchards. *Plants* **2022**, *11*, 2876. [\[CrossRef\]](#)
- Singerman, A.; Rogers, M.E. The economic challenges of dealing with citrus greening: The case of Florida. *J. Integr. Pest Manag.* **2020**, *11*, 3. [\[CrossRef\]](#)
- Wang, S.; Xie, W.; Yan, X. Effects of future climate change on citrus quality and yield in China. *Sustainability* **2022**, *14*, 9366. [\[CrossRef\]](#)
- Huang, Z.; Li, Z.; Yao, L.; Yuan, Y.; Hong, Z.; Huang, S.; Wang, Y.; Ye, J.; Zhang, L.; Ding, J. Geographical distribution and potential distribution prediction of thirteen species of Citrus L. in China. *Environ. Sci. Pollut. Res.* **2024**, *31*, 6558–6571. [\[CrossRef\]](#) [\[PubMed\]](#)
- Liang, Y.; Jiang, W.; Liu, Y.; Wu, Z.; Zheng, R. Picking-Point Localization Algorithm for Citrus Fruits Based on Improved YOLOv8 Model. *Agriculture* **2025**, *15*, 237. [\[CrossRef\]](#)
- Xiao, X.; Wang, Y.N.; Jiang, Y.M. End-Effectors Developed for Citrus and Other Spherical Crops. *Appl. Sci.* **2022**, *12*, 7945. [\[CrossRef\]](#)
- Chen, Z.Q.; Lei, X.H.; Yuan, Q.C.; Qi, Y.N.; Ma, Z.B.; Qian, S.C.; Lyu, X. Key Technologies for Autonomous Fruit- and Vegetable-Picking Robots: A Review. *Agronomy* **2024**, *14*, 2233. [\[CrossRef\]](#)
- Lu, J.; Lee, W.S.; Gan, H.; Hu, X. Immature citrus fruit detection based on local binary pattern feature and hierarchical contour analysis. *Biosyst. Eng.* **2018**, *171*, 78–90. [\[CrossRef\]](#)

9. Wu, G.; Li, B.; Zhu, Q.; Huang, M.; Guo, Y. Using color and 3D geometry features to segment fruit point cloud and improve fruit recognition accuracy. *Comput. Electron. Agric.* **2020**, *174*, 105475. [[CrossRef](#)]
10. Dubey, S.R.; Jalal, A.S. Apple disease classification using color, texture and shape features from images. *Signal Image Video Process.* **2016**, *10*, 819–826. [[CrossRef](#)]
11. Farooque, A.A.; Hussain, N.; Schumann, A.W.; Abbas, F.; Afzaal, H.; McKenzie-Gopsill, A.; Esau, T.; Zaman, Q.; Wang, X. Field evaluation of a deep learning-based smart variable-rate sprayer for targeted application of agrochemicals. *Smart Agric. Technol.* **2023**, *3*, 100073. [[CrossRef](#)]
12. Wu, H.R.; Li, X.X.; Sun, F.C.; Huang, L.M.; Yang, T.; Bian, Y.C.; Lv, Q.R. An Improved Product Defect Detection Method Combining Centroid Distance and Textural Information. *Electronics* **2024**, *13*, 3798. [[CrossRef](#)]
13. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
14. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
15. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [[CrossRef](#)] [[PubMed](#)]
16. Mai, X.; Zhang, H.; Jia, X.; Meng, M.Q.-H. Faster R-CNN with classifier fusion for automatic detection of small fruits. *IEEE Trans. Autom. Sci. Eng.* **2020**, *17*, 1555–1569. [[CrossRef](#)]
17. Fu, L.; Majeed, Y.; Zhang, X.; Karkee, M.; Zhang, Q. Faster R-CNN-based apple detection in dense-foliage fruiting-wall trees using RGB and depth features for robotic harvesting. *Biosyst. Eng.* **2020**, *197*, 245–256. [[CrossRef](#)]
18. Lu, J.; Yang, R.; Yu, C.; Lin, J.; Chen, W.; Wu, H.; Chen, X.; Lan, Y.; Wang, W. Citrus green fruit detection via improved feature network extraction. *Front. Plant Sci.* **2022**, *13*, 946154. [[CrossRef](#)] [[PubMed](#)]
19. Hussain, M. YOLOv5, yolov8 and yolov10: The go-to detectors for real-time vision. *arXiv* **2024**, arXiv:2407.02988.
20. 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 Computer Vision–ECCV 2016: 14th European Conference, Part I 14, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
21. Redmon, J. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
22. Gu, B.; Wen, C.J.; Liu, X.Z.; Hou, Y.J.; Hu, Y.H.; Su, H.Q. Improved YOLOv7-Tiny Complex Environment Citrus Detection Based on Lightweighting. *Agronomy* **2023**, *13*, 2667. [[CrossRef](#)]
23. Li, G.J.; Huang, X.J.; Ai, J.Y.; Yi, Z.R.; Xie, W. Lemon-YOLO: An efficient object detection method for lemons in the natural environment. *IET Image Process.* **2021**, *15*, 1998–2009. [[CrossRef](#)]
24. Ou, J.J.; Zhang, R.H.; Li, X.M.; Lin, G.C. Research and Explainable Analysis of a Real-Time Passion Fruit Detection Model Based on FSOne-YOLOv7. *Agronomy* **2023**, *13*, 1993. [[CrossRef](#)]
25. Xu, L.J.; Wang, Y.H.; Shi, X.S.; Tang, Z.L.; Chen, X.Y.; Wang, Y.C.; Zou, Z.Y.; Huang, P.; Liu, B.; Yang, N.; et al. Real-time and accurate detection of citrus in complex scenes based on HPL-YOLOv4. *Comput. Electron. Agric.* **2023**, *205*, 107590. [[CrossRef](#)]
26. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
27. Sun, F.; Lv, Q.; Bian, Y.; He, R.; Lv, D.; Gao, L.; Wu, H.; Li, X. Grape Target Detection Method in Orchard Environment Based on Improved YOLOv7. *Agronomy* **2025**, *15*, 42. [[CrossRef](#)]
28. Ma, C.J.; Fu, Y.Y.; Wang, D.Y.; Guo, R.; Zhao, X.Y.; Fang, J. YOLO-UAV: Object Detection Method of Unmanned Aerial Vehicle Imagery Based on Efficient Multi-Scale Feature Fusion. *IEEE Access* **2023**, *11*, 126857–126878. [[CrossRef](#)]
29. Chen, J.; Kao, S.-h.; He, H.; Zhuo, W.; Wen, S.; Lee, C.-H.; Chan, S.-H.G. Run, don't walk: Chasing higher FLOPS for faster neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition, 2023, Vancouver, BC, Canada, 17–24 June 2023; pp. 12021–12031.
30. Tang, Z.X.; Zhang, W.; Li, J.L.; Liu, R.; Xu, Y.S.; Chen, S.Y.; Fang, Z.Y.; Zhao, F.C.L. LTSCD-YOLO: A Lightweight Algorithm for Detecting Typical Satellite Components Based on Improved YOLOv8. *Remote Sens.* **2024**, *16*, 3101. [[CrossRef](#)]
31. Zhu, L.; Wang, X.; Ke, Z.; Zhang, W.; Lau, R.W. Biformer: Vision transformer with bi-level routing attention. In Proceedings of the IEEE/CVF Conference On Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 10323–10333.
32. Sun, Y.; Li, Y.; Li, S.; Duan, Z.H.; Ning, H.A.; Zhang, Y.H. PBA-YOLOv7: An Object Detection Method Based on an Improved YOLOv7 Network. *Appl. Sci.* **2023**, *13*, 10436. [[CrossRef](#)]
33. Qi, Y.; He, Y.; Qi, X.; Zhang, Y.; Yang, G. Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 6070–6079.

34. Li, H.; Li, J.; Wei, H.; Liu, Z.; Zhan, Z.; Ren, Q. Slim-neck by GSConv: A better design paradigm of detector architectures for autonomous vehicles. *arXiv* **2022**, arXiv:2206.02424.
35. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 12993–13000.
36. Ma, S.; Xu, Y. Mpdious: A loss for efficient and accurate bounding box regression. *arXiv* **2023**, arXiv:2307.07662.
37. Cheng, D.G.; Zhao, Z.Q.; Feng, J. Rice Diseases Identification Method Based on Improved YOLOv7-Tiny. *Agriculture* **2024**, *14*, 709. [[CrossRef](#)]
38. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
39. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
40. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
41. Yang, L.; Zhang, R.-Y.; Li, L.; Xie, X. Simam: A simple, parameter-free attention module for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 11863–11874.
42. Zhang, Y.-F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* **2022**, *506*, 146–157. [[CrossRef](#)]
43. Gevorgyan, Z. SIoU loss: More powerful learning for bounding box regression. *arXiv* **2022**, arXiv:2205.12740.
44. Malta, A.; Mendes, M.; Farinha, T. Augmented reality maintenance assistant using yolov5. *Appl. Sci.* **2021**, *11*, 4758. [[CrossRef](#)]
45. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
46. Varghese, R.; Sambath, M. YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness. In Proceedings of the 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), Chennai, India, 18–19 April 2024; pp. 1–6.
47. Zheng, Z.; Xiong, J.; Lin, H.; Han, Y.; Sun, B.; Xie, Z.; Yang, Z.; Wang, C. A method of green citrus detection in natural environments using a deep convolutional neural network. *Front. Plant Sci.* **2021**, *12*, 705737. [[CrossRef](#)] [[PubMed](#)]
48. Lin, Y.; Huang, Z.; Liang, Y.; Liu, Y.; Jiang, W. Ag-yolo: A rapid citrus fruit detection algorithm with global context fusion. *Agriculture* **2024**, *14*, 114. [[CrossRef](#)]
49. Deng, F.; Chen, J.; Fu, L.; Zhong, J.; Qiao, W.; Luo, J.; Li, J.; Li, N. Real-time citrus variety detection in orchards based on complex scenarios of improved YOLOv7. *Front. Plant Sci.* **2024**, *15*, 1381694. [[CrossRef](#)]
50. Liao, Y.; Li, L.; Xiao, H.; Xu, F.; Shan, B.; Yin, H. YOLO-MECD: Citrus Detection Algorithm Based on YOLOv11. *Agronomy* **2025**, *15*, 687. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.