applied sciences

MDPI

*Article*

# Target Detection for Coloring and Ripening Potted Dwarf Apple Fruits Based on Improved YOLOv7-RSES

**Haoran Ma [1], Yanwen Li [2], Xiaoying Zhang [1,*], Yaoyu Li [3], Zhenqi Li [1], Runqing Zhang [1], Qian Zhao [1] and Renjie Hao [4]**

[1] College of Software, Shanxi Agricultural University, Jinzhong 030801, China; mhrsxau@163.com (H.M.); z20213628@stu.sxau.edu.cn (Z.L.); zrq15611636180@163.com (R.Z.); 15525052255@163.com (Q.Z.)
[2] College of Information Science and Engineering, Shanxi Agricultural University, Taigu 030801, China; lyw@sxau.edu.cn
[3] School of Agricultural Engineering, Shanxi Agricultural University, Taigu 030801, China; 18406550256@163.com
[4] College of Resources and Environment, Shanxi Agricultural University, Taigu 030801, China; 17634099957@163.com
* Correspondence: xiaoyingzhang@sxau.edu.cn; Tel.: +86-158-0344-9361

**Abstract:** Dwarf apple is one of the most important forms of garden economy, which has become a new engine for rural revitalization. The effective detection of coloring and ripening apples in complex environments is important for the sustainable development of smart agricultural operations. Addressing the issues of low detection efficiency in the greenhouse and the challenges associated with deploying complex target detection algorithms on low-cost equipment, we propose an enhanced lightweight model rooted in YOLOv7. Firstly, we enhance the model training performance by incorporating the Squeeze-and-Excite attention mechanism, which can enhance feature extraction capability. Then, an SCYLLA-IoU (SIoU) loss function is introduced to improve the ability of extracting occluded objects in complex environments. Finally, the model was simplified by introducing depthwise separable convolution and adding a ghost module after up-sampling layers. The improved YOLOv7 model has the highest AP value, which is 10.00%, 5.61%, and 6.00% higher compared to YOLOv5, YOLOv7, and YOLOX, respectively. The improved YOLOv7 model has an MAP value of 95.65%, which provides higher apple detection accuracy compared to other detection models and is suitable for potted dwarf anvil apple identification and detection.

**Keywords:** potted dwarf rootstock apple; YOLOv7; Squeeze–Excite; SCYLLA-IoU



**Citation:** Ma, H.; Li, Y.; Zhang, X.; Li, Y.; Li, Z.; Zhang, R.; Zhao, Q.; Hao, R. Target Detection for Coloring and Ripening Potted Dwarf Apple Fruits Based on Improved YOLOv7-RSES. *Appl. Sci.* **2024**, *14*, 4523. https://doi.org/10.3390/app14114523

Academic Editors: Salik Khanal and Vitor Filipe

Received: 23 April 2024
Revised: 18 May 2024
Accepted: 20 May 2024
Published: 24 May 2024



## 1. Introduction

Potting cultivated dwarfing rootstock apple [1] is an innovative approach to apple cultivation. It involves reducing the tree size [2–4], promoting early fruiting, increasing the yield, and facilitating the renewal and management of apple varieties [5]. This method not only enhances the efficiency of natural resources such as land and water but also preserves resources. It provides a new avenue for improving the economic benefits of orchards and promoting the development of rural industries. Concurrently, it fosters the growth of the courtyard economy, which emphasizes ecological balance and environmental protection. Through rational planting and breeding patterns, it can improve soil quality, maintain clean water sources, and enhance biodiversity. Additionally, the courtyard economy can promote the recycling of waste, thereby reducing environmental pollution. In agricultural production, real-time and accurate monitoring of fruit quantity, automatic harvesting [6,7], and precise yield prediction are beneficial for improving production efficiency and economic benefits [8]. Moreover, they help to drive the innovation and advancement of agricultural technology, promoting the sustainable development of agriculture. Consequently, detecting

the growth conditions of potted dwarfing rootstock apple trees in an effective way holds substantial practical significance for advancing precision agriculture [9].

Apple varieties are typically characterized by attributes such as color, shape, and texture. In China, where apples are a significant horticultural crop, the adoption of machine learning (ML) and computer vision (CV) technologies has advanced the capabilities for apple detection and recognition. This progress supports the intelligent production of apples [10]. For example, Guo et al. proposed a detection algorithm based on image color region segmentation [11]. This algorithm assesses an object's eligibility as a detection target by analyzing its different features. However, traditional algorithms frequently encounter challenges in practical applications due to factors such as foliage occlusion, variable lighting conditions, and fruit entanglement. The boundaries of the corn stover section were effectively extracted using a shape feature-based image segmentation algorithm [12].

In recent years, the YOLO algorithm has been widely adopted in precision agriculture, significantly improving the speed and accuracy of crop detection [13,14]. YOLOv7, an advanced deep learning algorithm based on machine vision, effectively addresses the challenges of object detection [15]. For example, the improved YOLOv7 has been utilized for the recognition of apple inflorescence morphology [16]. Experimental results indicate that the improved YOLOv7 model achieves a recognition speed of 42.58 frames per second (fps), outperforming other models such as YOLOv5s, the improved YOLOv5, and the original YOLOv7. In [17], the YOLOv7 model was combined with a Squeeze–Excite (SE) network and performed well in a detection task on the PASCAL VOC dataset. The improved model reduced the number of parameters by 12.3% and the FLOPs by 18.86% compared to the original YOLOv7. Furthermore, the advancements in YOLOv3 based on SIoU loss [18] have led to improved accuracy in detecting occluded pedestrians. The YOLOv7 model has demonstrated significant theoretical and practical applications in the detection of potted dwarf apple trees.

In this paper, the tree characteristics of potted dwarf anvil apple trees are taken as the subject of the research, emphasizing and analyzing the two key periods for the automated management and yield prediction of fruit trees: the coloring and ripening periods. Fruit farmers and orchard managers can monitor the number and condition of fruits undergoing coloration in real time. This allows them to promptly implement management measures, such as adjusting light exposure, controlling moisture, and applying fertilizers, to optimize the coloration effect of the fruits. By analyzing the images of these two specific periods with deep learning, this paper aims to improve the recognition accuracy to better serve the automated management and yield prediction of potted dwarf anvil apples, thereby reducing environmental pollution and realizing green and eco-friendly orchard production.

The lightweight operation of the YOLOv7 model was ensured by adopting an efficient and parameterizable backbone network called RepBlock, which ensures accuracy while preserving computation time. Furthermore, it introduces the SE attention mechanism and the SIoU loss function, which further reduce complexity with minimal accuracy loss, enhancing the model's feature extraction capability in complex environments and thus improving the detection of obscured objects. This paper focuses on the coloring and ripening process of potted dwarf rootstock apple trees and explores the utilization of tree features to enhance the accuracy of fruit tree identification and detection. The proposed model is designed to enhance the accuracy of recognizing and detecting potted dwarf anvil apples, providing effective technical support for smart agriculture automated picking equipment.

## 2. Materials and Methods

### *2.1. Dataset Collection*

This work focuses on researching the new variety of Ruixiang red apple with dwarf anvil, using pot planting in a sunlit greenhouse. The sunlit greenhouse spans 4000 square meters and houses 1500 potted dwarf anvil fruit trees. The rows are 3 m apart, with trees spaced 2 m apart within each row. The difference between dwarf rootstock fruit trees and traditionally planted fruit trees can be visually compared through Figure 1. The dataset

was collected at the fruit tree dwarf rootstock planting base of Shanxi Agricultural University, located in Xiaowang Village, Nandali Township, Xia County, Yuncheng City, Shanxi Province, with geographic coordinates ranging from 111°02′ to 111°41′ east longitude and 34°55′ to 35°19′ north latitude. The collection period was from September 15 to 30 October 2023, using a Canon 700D camera. The images were captured in various weather and light conditions (sunny or cloudy) and at different distances—far (100–150 cm) and close (50–100 cm)—as shown in Figure 2. The apple fruit's reproductive period includes flowering, fruit drop, expansion, coloring, and ripening stages. This collection focuses on the coloring and ripening stage of the fruit. It is designed for fruits freshly taken out of the film bag, transitioning from green to yellow to red as they ripen. The fruit should have a diameter between 70 and 100 mm. During orchard production, it is crucial to monitor fruit color and ripeness in real time, automate picking, and accurately predict yield to boost efficiency and profits. Yuncheng, the research area, experiences year-round monsoon influence and has a warm temperate continental climate. It receives an average of 525 mm rainfall annually, enjoys 2350 h of sunshine, maintains an average yearly temperature of 13.3 °C, and has a frost-free period lasting 212 days. In recent years, through robust agricultural industry restructuring, the apple cultivation area has surged to 2.12 million mu, yielding over 1.8 billion kilograms annually, establishing itself as a prime national hub for high-quality apple production.

**Figure 1.** Potted dwarf apple trees and cultivated dwarf apple trees.

(**a**) (**b**) (**c**) (**d**)

**Figure 2.** *Cont.*

(**e**) (**f**) (**g**) (**h**)

**Figure 2.** Images of apples in different complex scenes. (**a**) Fairing; (**b**) backlight; (**c**) coloring period; (**d**) maturity; (**e**) close proximity; (**f**) long distance; (**g**) blade occlusion; (**h**) branch occlusion.

This paper used an on-site data collection approach to tackle uneven sample distribution, enhance dataset diversity, and reduce bias in model training. We focused on target detection technology for tasks like machine inspection and mechanical fruit picking, where identifying ripe fruits at different distances and executing picking operations in close proximity are crucial. We gathered apple images from various time slots, distances, and angles, totaling 1325 images.

### *2.2. Dataset Production*

In this paper, we utilized the LabelImg software (1.8.6) for manual annotation of the images. We excluded apple images with over two-thirds occlusion to ensure annotation accuracy. This decision considers both apple appearance and shape and the picking robot end-effector's fault tolerance. After completing the annotation, an XML file was generated containing category and coordinate information. We divided the dataset into training, validation, and test sets using an 8:1:1 ratio. The test set comprises 150 images, including 50 showcasing apples in various complex scenes, and serves to evaluate the model's detection efficacy. Figure 2 presents sample images of apples in diverse complex scenarios. The sufficiency of image data is pivotal during the model training phase. Insufficient training data may result in overfitting. Presently, we possess 1325 original apple images, which partly capture apple characteristics. However, they may not fully cover the differences in light, weather, noise, clarity, and other factors found in natural environments. Hence, to enhance the performance and generalization capability of the target detection model, it is imperative to expand the apple image dataset. We took several steps to enhance our data, resulting in 7950 images, as shown in Figure 3. This expansion is crucial for enhancing the model's accuracy and generalization ability.

$$I_{\text{new}} = aI_{old} + b \tag{1}$$

$I_{old}$ is the original image, $I_{new}$ is the preprocessed image, and $a$ and $b$ are the adjustment parameters. The expanded image dataset will offer richer and more detailed target features for subsequent studies, thereby enhancing the model's capability to discern fruit shapes and features.

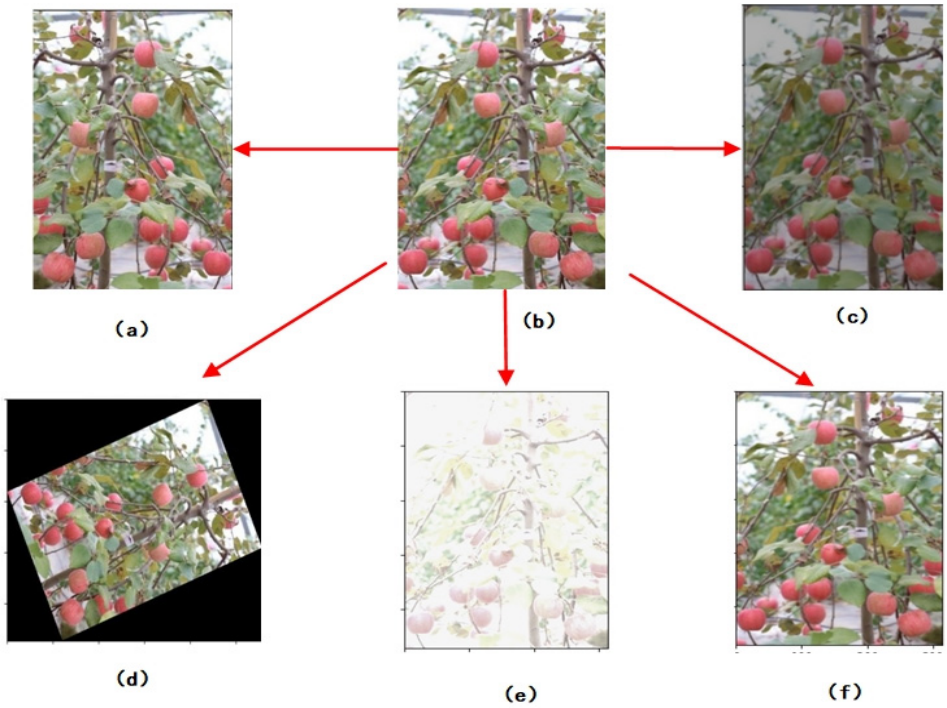


**Figure 3.** Example of image enhancement. (**a**) Random contrast; (**b**) artwork; (**c**) random brightness; (**d**) random rotation; (**e**) random noise; (**f**) random flipping.

### *2.3. YOLOv7 Principle and Structure*

The YOLOv7 model has three main parts: the backbone feature extraction network, the neck network, and the detection head network [19], as depicted in Figure 4. The backbone processes the input image with 640 pixels × 640 pixels and extracts valuable information via feature extraction. This process mainly depends on the collaboration of the Conv+BN+Silu (CBS) module, the Efficient Aggregation Network (ELAN) module, the MaxPool (MP) module, and the SPPCSPC module [20].

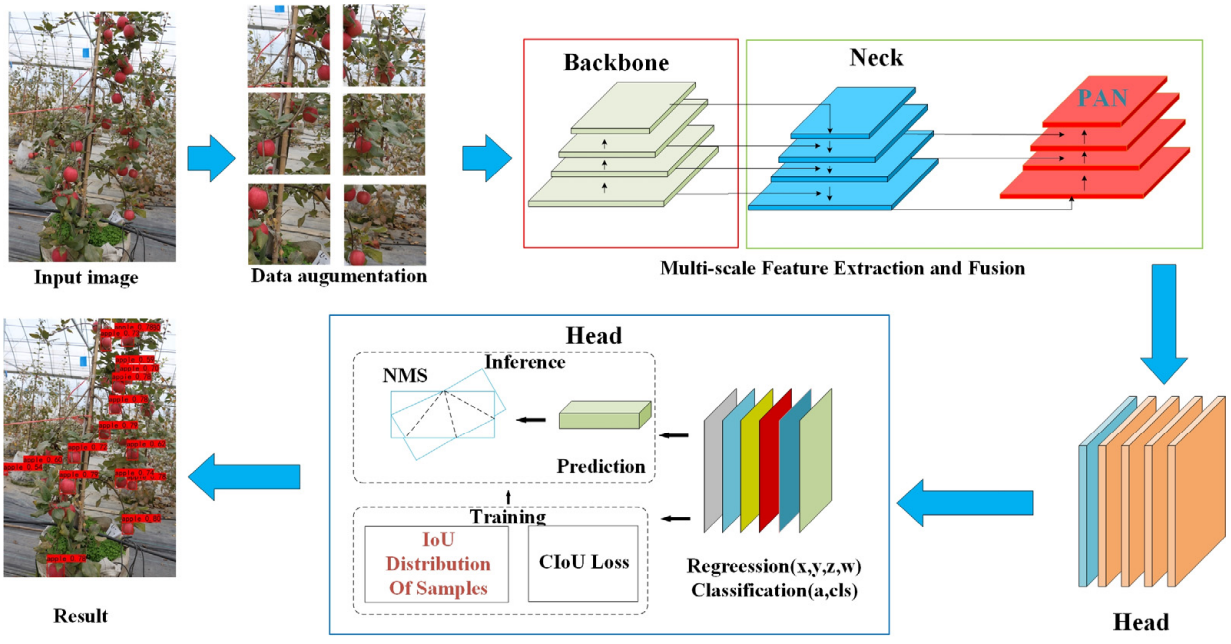


**Figure 4.** Network architecture of YOLOv7.

In the backbone, the Multi-Concat-Block and Transition-Block play pivotal roles. The Multi-Concat-Block comprises four branches, each conducting a varying number of normalized convolution operations on the input feature layer. The results from these four branches are combined and then processed through another normalization convolution operation, producing a final output feature layer of fixed size. This design boosts network depth to enhance accuracy while tackling gradient vanishing issues with its multi-branch stacking and skip connection structure.

The neck network, on the other hand, uses the PANet network structure [21,22]. This structure combines top-down and bottom-up information flows through a two-way fusion strategy. The neck consolidates information from different backbone and detection layers, enabling multi-scale fusion and extracting 3 enhanced feature layers.

Finally, the detection head network carefully examines feature points in each layer to determine the exact location, confidence level, and category of the target object.

YOLOv7 builds a feature pyramid structure like YOLOv5 on the three key feature layers of the backbone network to boost feature extraction. This architecture utilizes Multi-Concat-Block and Transition-Block modules to combine feature layers of varying scales through up-sampling and down-sampling operations, enhancing the extraction of high-quality feature layers for target detection. It is worth noting that YOLOv7 utilizes the unique SPPCSPC module in the early stage to perform deep feature extraction for feature layers of size (20,20,1024), aiming at expanding the sensory field of the network layer. The SPPCSPC module combines the CSP and SPP modules. It divides the feature layer into two parts: one passes through the module after standardized convolution, while the other undergoes normalized convolution. Additionally, the second part is pooled using the SPP module with four different kernel sizes to integrate features into a fixed-size output layer. This not only helps to improve the sensory field and accuracy of the network layer, but also effectively circumvents the repeated extraction of feature information, reduces the computational complexity, and thus improves the computational speed.

### *2.4. Improvements to YOLOv7*

The network architecture of YOLOv7 consists of three main parts: the input layer, the backbone network, and the detection head [19]. On the input side, YOLOv7 follows the Mosaic data enhancement approach proposed by YOLOv4, which is trained by randomly cropping four images and splicing them into a single image, thus enriching the dataset and improving the training efficiency, while keeping the training and inference costs unchanged.

To strike a balance between detection accuracy and efficiency, YOLOv7 introduces a series of innovative strategies, including Reparametrized Convolution (RepConv), the Efficient Lightweight Aggregation Network (ELAN), and dynamic label assignment. However, it is noteworthy that YOLOv7 primarily relies on convolutional operations for feature extraction, which are inherently local operations. Convolutional layers often only model relationships between neighboring pixels, making it difficult to capture long-range dependencies. This limitation may result in the loss of features for small objects and missed detections for immature fruits. To address this issue, an attention mechanism is introduced to pay closer attention to the relationships between pixel features. Serving as a global operation, the attention mechanism computes weights between features through matrix operations, enabling the model to better capture long-range dependencies. The image features obtained from self-attention mechanisms complement those obtained from convolutional operations, collectively enhancing the model's performance.

In addition, the originally utilized CIoU loss function in YOLOv7 only accounts for the scale loss of the bounding boxes, without considering the mismatch between the predicted and true box orientations. Therefore, this paper adopts the SIoU loss function to replace the CIoU loss function, incorporating orientation loss into the model training process, aiming to further enhance the model's performance.

2.4.1. SE Attention Mechanism

Traditional CNNs emphasize feature representation within channels while overlooking the mapping relationships between channels. The SE module addresses inter-channel relationships through squeeze-and-excitation operations, thereby adaptively adjusting channel responses. In traditional convolutional networks, inter-channel relationships are often implicit and are confined to specific hierarchical levels. However, top-level channels are closely related to tasks; for instance, in segmentation networks, the number of top-level channels corresponds to the number of segmentation categories. In middle layers of the network, the number of channels is typically based on empirical data or test results, which may be derived from real-world applications or extensive experimental analysis. The SE network architecture, proposed in [23], integrates the idea of attention mechanisms into convolutional neural networks to achieve adaptive learning of the importance of each channel.

The SE module introduces a parameter-efficient channel attention mechanism, as illustrated in Figure 5. It achieves this through two steps: squeezing and excitation. In the squeezing phase, the feature maps are aggregated into a scalar value, and in the excitation phase, a fully connected layer is utilized to transform this value into a weight vector, which is then used to weight the feature maps. The SE module enables CNNs to learn the importance of each channel, thereby improving model performance. Demonstrating strong performance across various image classification tasks, the SE network has been widely adopted in diverse visual tasks.

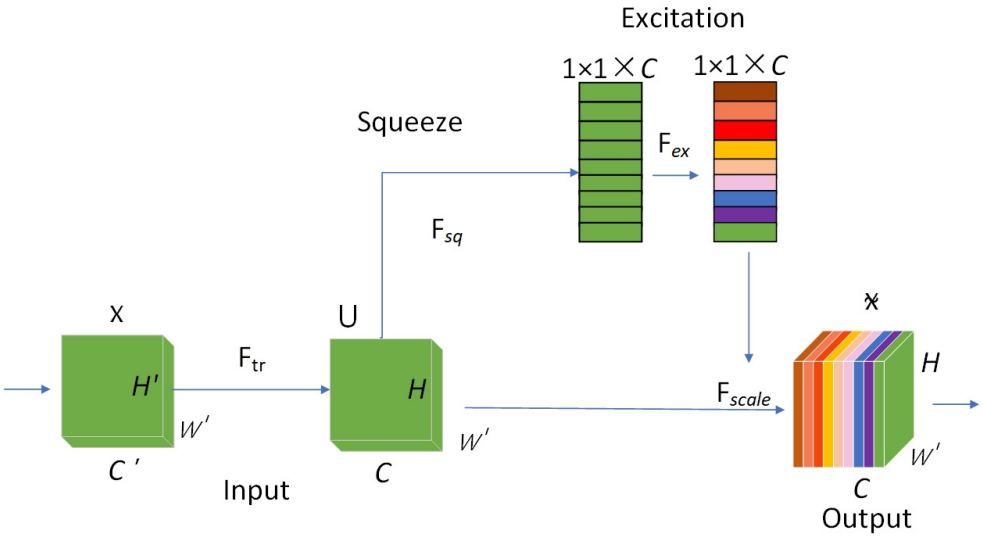


**Figure 5.** SE attention mechanism.

The structure of the SE block is to map the input feature map $X \in R^{H' \times W' \times C'}$ to the output feature map $U \in R^{H \times W \times C}$ through convolutional operation Ftr, where the collection of convolutional kernels is represented by $V = [v_1, v_2, \ldots, v_c]$I, and the output is represented by $U = [u_1, u_2, \ldots, u_c]$; thus,

$$u_c = u_c * X = \sum_{s=1}^{C'} v_c^s * x^s \tag{2}$$

This formula represents the convolution operation, where $u_c \in R^{H \times W}$, $v_c = \left[v_c^1, v_c^2, \ldots, v_c^{C'}\right]$, and $X = \left[x^1, x^2, \ldots, x^{C'}\right]$; $v_c^s$ is a 2D convolution kernel, indicating that $v_c$ acts on the corresponding channel of the $X$ squeeze stage, Global Information Integration, in order to comprehensively consider the information of each channel in the output feature map; this

paper adopts the global average pooling method to integrate global spatial information into a single-channel descriptor, $Z_c$

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \tag{3}$$

For activation and adaptive recalibration, in order to adequately capture the interdependencies between channels, this paper employs a simple gating mechanism, the sigmoid activation function, to achieve adaptive recalibration of channel responses.

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)) \tag{4}$$

where $\delta$ is the ReLU function, $W_1 \in R^{\frac{C}{r} \times C}$, and $W_2 \in R^{C \times \frac{C}{r}}$.

The final output result of the SE code block is obtained by rescaling U by a scale factor, s:

$$x = F_{scale}(u_c, s_c) = s_c u_c \tag{5}$$

In this paper, the SE module serves as a plug-and-play component designed to address inter-channel dependencies, as illustrated in Figure 6. It utilizes global average pooling to compress global spatial information, i.e., performing the squeeze operation. To fully leverage the aggregated information from the squeeze operation, the proposed model employed the FC-ReLU-FC-sigmoid operation to capture channel dependencies.

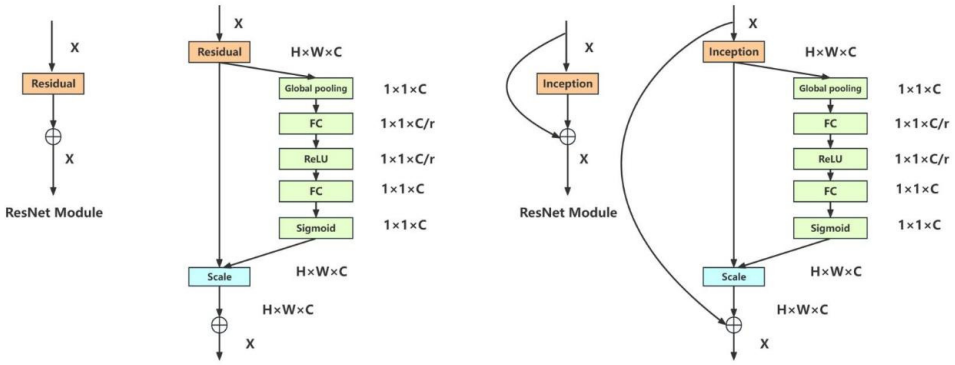


**Figure 6.** SE attention mechanism application.

The SE module can be added to any location of the YOLOv7 network, such as within the C3 module. When integrating the SE module into the YOLOv7 network via the common.py file, unlike the original C3 module, it incorporates the SE module into the bottom bottleneck section. This modification aims to facilitate better experimentation and evaluation of model performance.

#### 2.4.2. SIoU Loss Function

In the YOLOv7 algorithm, IoU is actually short for intersection over union, which is also known as the "intersection and union ratio". IoU has a crucial role in target detection and semantic segmentation. We can set the value of IoU as the ratio of the intersection and the union of two graph areas, as shown in Figure 7.

The predictor frame regression loss uses the CIoU loss, but CIoU does not take into account the mismatch between the predictor frame and the true frame directions. An SIoU loss function is introduced in which the penalty metrics are redefined by taking into account the angle of the vectors between the desired regressions. Applied to conventional neural networks and datasets, it is shown that SIoU improves the speed of training and the accuracy of inference. SIoU further considers the vector angles between the true and predicted frames, redefining the associated loss function. SIoU consists of four components—angular loss ($\Lambda$) (Angle Cost), distance loss ($\Delta$) (Distance Cost), shape loss ($\Omega$) (Shape Cost), and the

intersection and merger ratio loss (U) (IoU Cost)—as shown in the angular loss parameter schematic in Figure 8.

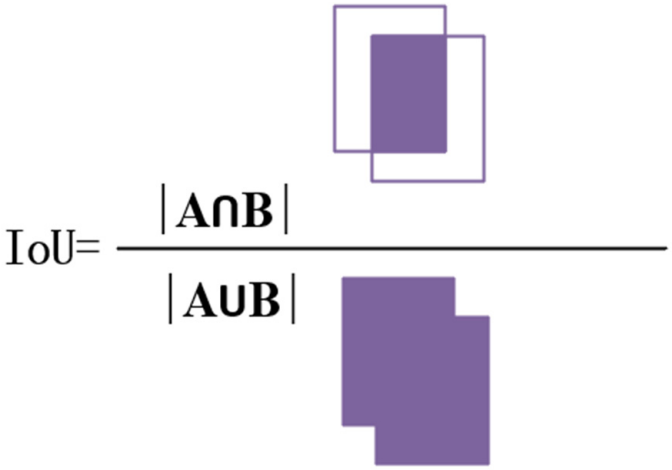


**Figure 7.** IoU interaction ratio.

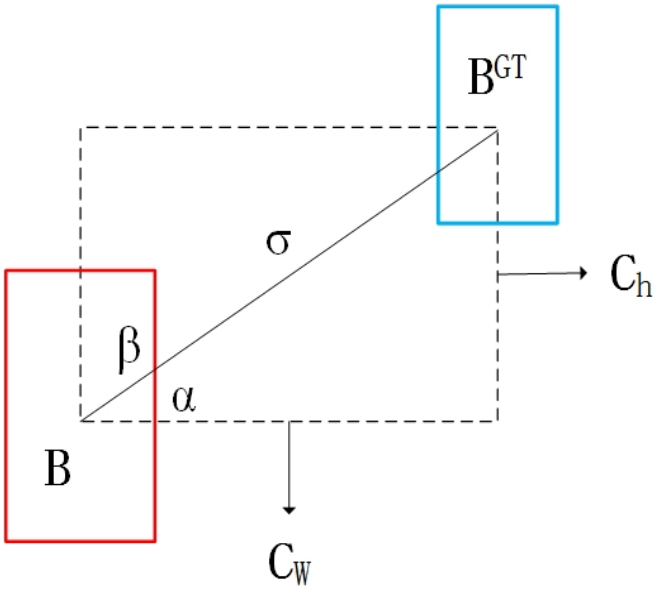


**Figure 8.** SIoU parameters.

IoU represents the intersection over union between the predicted bounding box (B) and the center of the real frame (BGT); Δ signifies the distance loss, aiming to minimize the centroid distances between various predicted bounding boxes; and Ω indicates the shape loss, which quantifies the deviation of the predicted bounding box's centroid from that of the center of the real frame.

If $\alpha \leq \pi/4$, the convergence process will first minimize $\alpha$ and otherwise minimize $\beta$:

$$\beta = \frac{\Omega}{2} - \alpha \tag{6}$$

The angular loss ($\Lambda$) is calculated as follows [13]:

$$\Lambda = 1 - 2\sin^2\left(\arcsin(x) - \frac{\pi}{4}\right) \tag{7}$$

$$x = \frac{c_h}{\sigma} = \sin(\alpha) \tag{8}$$

The distance loss (Δ) is calculated as follows:

$$\Delta = \sum_{t=x,y} (1 - e^{-\gamma p_t}) \tag{9}$$

Shape loss, Ω, is defined as follows:

$$\Omega = \sum_{t=w,h} (1 - e^{-w_t})^{\theta}, \tag{10}$$

$$\omega_w = \frac{|w - w^{gt}|}{max(w, w^{gt})} w_h = \frac{|h - h^{gt}|}{max(h, h^{gt})} \tag{11}$$

IoU loss is defined as follows:

$$L_{siou} = 1 - U + \frac{\Delta + \Omega}{2} \tag{12}$$

The total loss function is as follows:

$$L = W_{box} L_{box} + W_{cls} L_{cls} \tag{13}$$

2.4.3. Lightweight Improvement

RepBlock is obtained via structural re-parameterization; multi-branch networks often outperform single-path networks in classification tasks, yet this advantage can result in higher inference latency [24]. Our paper draws inspiration from RepVGG. We aim to balance accuracy and speed by adopting RepBlock, an efficient and tunable backbone network. RepBlock capitalizes on specific hardware acceleration features, and post-training, we significantly reduce inference latency by transforming the multi-branch topology into a single 3 × 3 convolution layer (RepConv) with ReLU activation. Figure 9 illustrates the backbone network structure of RepBlock.

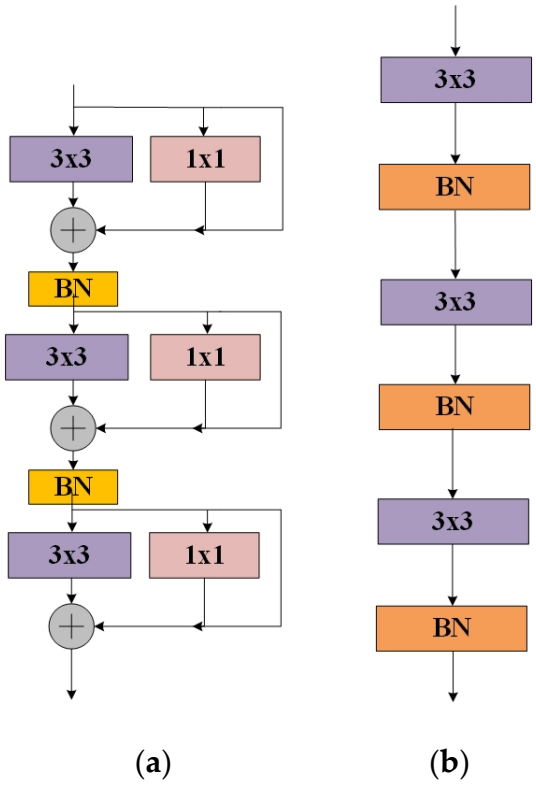


**Figure 9.** RepBlock structure. (**a**) Training structure; (**b**) inference structure.

In this paper, we performed an equivalent mapping on the original multi-branch architecture. Specifically, each branch had its convolutional kernels transformed into $3 \times 3$ convolutional kernels. For $1 \times 1$ convolutional layers, to address the shortcut branch problem without convolutional layers, we introduced four fixed numerical convolutional kernels, $K_1, K_2, K_3, K_4$, where $K_1, K_4$ are $3 \times 3$ convolution kernels with a central parameter of one and zero values around the center, and $K_2, K_3$ are $3 \times 3$ convolutional kernels with a center parameter of one and zero values around the center, while B and D are $3 \times 3$ convolutional kernels with all zero values [25]. It is worth noting that the introduction of convolutional kernels did not change the nature of the original convolutional operations. Figure 10 demonstrates the effectiveness of the equivalent mapping.

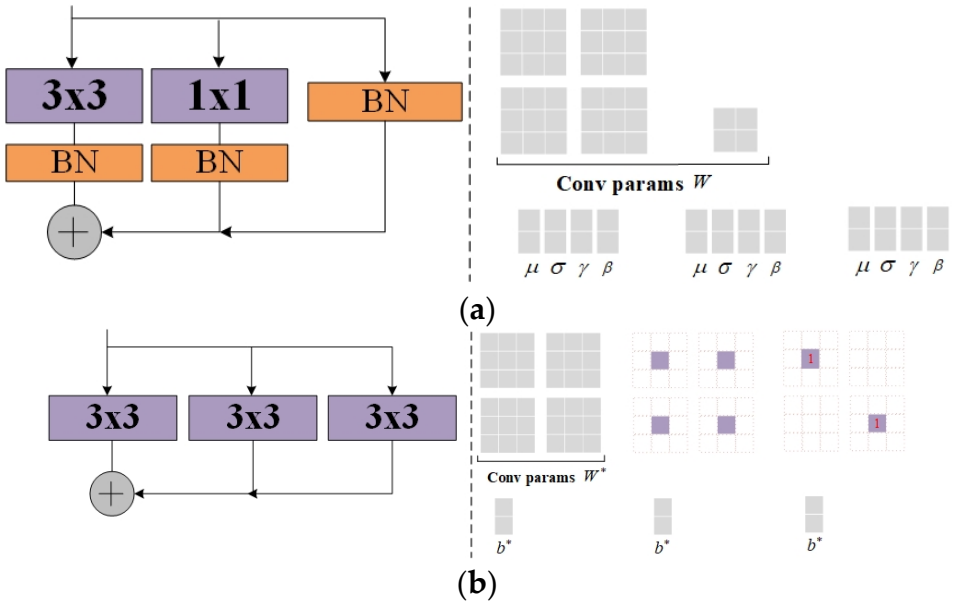


**Figure 10.** RepBlock equivalence mapping. (**a**) The structure and parameters of RepBlock; (**b**) the structure and parameters of RepBlock after equivalence mapping.

After equivalence mapping, we merged the convolutional layers and BN layers of each branch, resulting in a biased convolutional layer as shown in Figure 7. Specifically, we merged each branch into a single convolutional layer. The convolutional parameters and biases of the merged layer are denoted as $W_i, i = 1, 2, 3$ and $B_i, i = 1, 2, 3$, respectively. The structure of RepBlock is illustrated in Figure 8, and the relationship between its input and output operations is represented by Equation (14).

$$\begin{aligned} Y &= (X * W_1 + b_1) + (x * W_2 + b_2) + (x * W_3 + b_3) \\ &= I * (W_1 + K_2 + W_3) + (b_1 + b_2 + b_3) \end{aligned} \tag{14}$$

Then, letting $W^* = W_1 + W_2 + W_3$ and $b^* = b_1 + b_2 + b_3$, substituting into Equation (14), we obtain Equation (15).

$$\begin{aligned} Y &= (X * K_1 + b_1) + (X * K_2 + b_2) + (X * K_3 + b_3) \\ &= X * W^* + b^* \end{aligned} \tag{15}$$

In this paper, we successfully utilized the convolutional kernel fusion technique described in Equation (15) to convert three sets of parameters $\{W_i, B_i, i = 1, 2, 3\}$ into a single set of parameters $\{W^*, B^*\}$, as shown in Figure 11. This method effectively reduced the number of convolutional kernels by two-thirds and significantly improved the inference efficiency of the network.

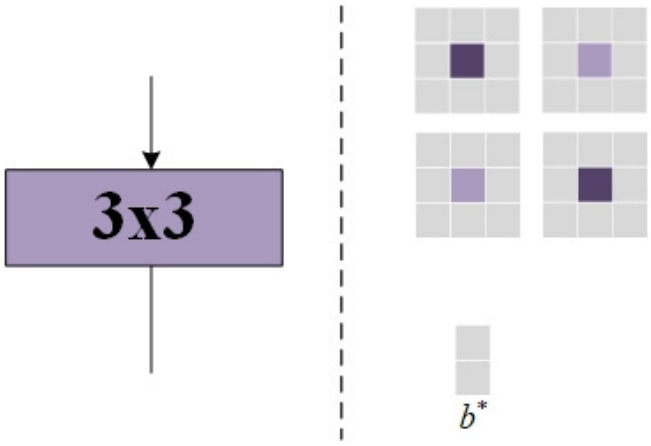


**Figure 11.** Multi-branch integration.

## 3. Experimental Design and Results Analysis

### 3.1. Experimental Environment

In this paper, the experimental tests were carried out on the deep learning server provided by the Smart Agriculture Laboratory of Shanxi Agricultural University. The server is powered by the Windows 10 operating system and is equipped with an Intel® Core™ i7-7700HQ processor (Intel, Santa Clara, CA, USA) (clocked at 3.80 GHz) and an NVIDIA GeForce GTX 3090 graphics card (NVIDIA, Santa Clara, CA, USA). The software used in the experiment and their versions are shown in Table 1.

**Table 1.** System environment of the experiment.

| Software | Version |
|---|---|
| CUDA | 9.2.4 |
| Python | 10.3.1 |
| Visual Studoo Code | 1.54.5 |
| Tensorflow-gpu | 1.14.3 |

### 3.2. Evaluation Indicators

To assess the detection performance of the model, this paper utilizes three metrics—precision (*P*), recall (*R*) and average precision (*AP*) (PK)—as the evaluation criteria to assess the model [21] (QH). Accuracy, P, is the ratio of true positive detections to the total number of positive predictions made by the model, and is calculated by the following formula [15]:

$$P = \frac{TP}{TP + FP} \times 100\% \tag{16}$$

*R* measures the ratio of true positive detections to the total number of actual positive instances. It is calculated as followed:

$$R = \frac{TP}{TP + FN} \times 100\% \tag{17}$$

where *TP* denotes the number of true positive detections, *FP* represents the number of false positive detections, and *FN* indicates the number of false negative detections.

The average precision (*AP*) represents the area under the precision–recall curve, which is plotted by combining points from both precision and recall metrics. The formula is

$$AP = \int_0^1 P(R)\mathrm{d}R \times 100\% \tag{18}$$

where $P(R)$ is the value of *P* corresponding to *R* on the PR curve [22].

### *3.3. Improved Experimental Precision*

The improved YOLOv7 network underwent systematic training on a training set containing 1340 images of potted dwarf apple trees. To verify the effectiveness of this method, we further evaluated 333 images of dwarf apple trees from an independent test set. According to the results presented in Table 2, the proposed model demonstrated excellent performance, with an accuracy rate of up to 87.22% and a recall rate of 90.02%. These results fully prove the reliability and practicality of the improved model. Figure 12 presents the test accuracy of the experimental models.

**Table 2.** Optimized model testing.

| Model | Precision (%) | Recall (%) |
|---|---|---|
| Improved YOLOv7 | 87.22 | 90.02 |

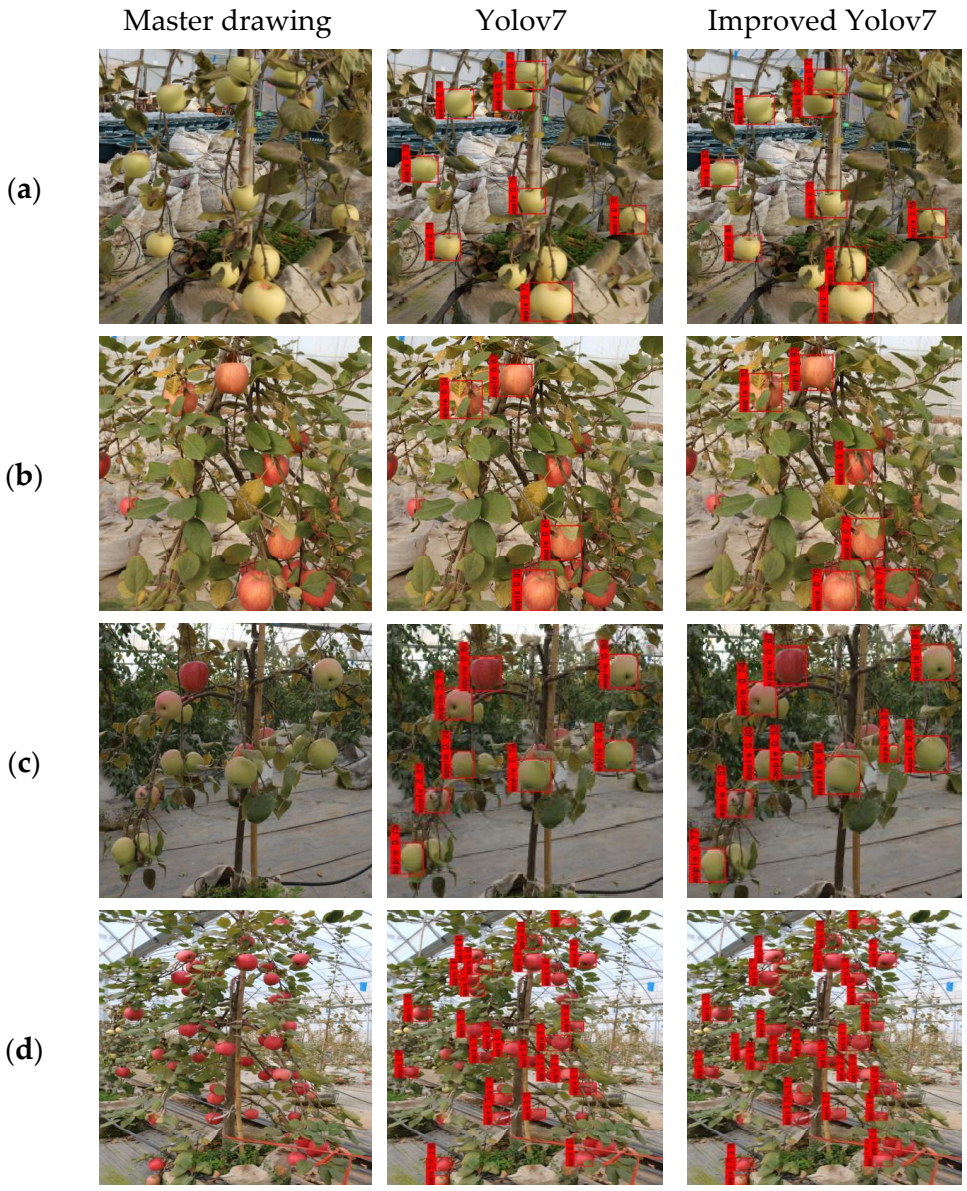


**Figure 12.** *Cont.*


**Figure 12.** YOLOv7: comparative analysis before and after improvement. (**a**) Coloring period; (**b**) maturity period; (**c**) backlight; (**d**) smooth light; (**e**) close distance; (**f**) long distance.

## 4. Discussion

To further validate the advantages and efficacy of the proposed improved YOLOv7 model for potted dwarf apple detection tasks in orchards, we meticulously adjusted the model parameters, and a series of comparative experiments were conducted. These experiments used the industry-representative target detection models, namely YOLOv7, YOLOv5, and YOLOX. The experimental results demonstrated the performance differences between the algorithms in the dwarf apple recognition detection task, as shown in Figure 13, by comparing the performance of YOLOX, YOLOv5, YOLOv7, and the improved YOLOv7 model in key metrics such as recognition accuracy, recall rate, and average precision.


**Figure 13.** *Cont.*

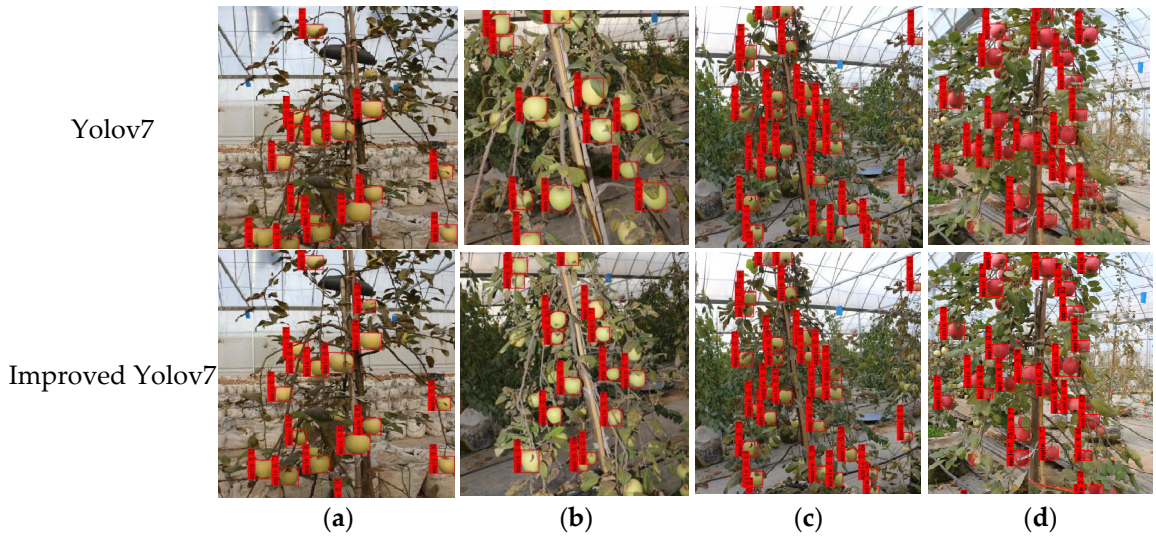


**Figure 13.** Detection results of different models with various conditions. (**a**) Near distance; (**b**) long distance; (**c**) coloring stage; (**d**) maturity stage.

After comparing the close-up detection effect images, we observed that all four detection models exhibited certain detection capabilities. Among them, YOLOv7 stood out in improving the precision of the bounding boxes, with both the completeness and accuracy of its bounding boxes surpassing the other models. All models showed varying degrees of deficiencies in the completeness of detection annotations.

In the long-range small target detection effect images, by introducing the SE attention mechanism, our method significantly enhanced the feature extraction capability. Under these circumstances, YOLOX and YOLOv5 failed to detect all targets completely, while YOLOv7 successfully detected all targets but made misjudgments in the recognition during the coloring and maturation periods. This indicates that the improved YOLOv7 not only effectively prevents the omission of small targets in long-range images but also enhances detection accuracy. Moreover, the improved algorithm demonstrated optimal detection performance when processing dense images. In summary, the improved YOLOv7 exhibited outstanding precise recognition capabilities in various environments, including long-range, close-up, and dense scenes.

As shown in Table 3, the detection accuracy of the YOLOX model for potted dwarf apples was 81.91%, the recall rate was 86.35%, and the average precision was 89.65%; the YOLOv5 model had a detection accuracy of 79.50%, a recall rate of 81.83%, and an average precision of 86.65%; the YOLOv7 model had a detection accuracy of 82.34%, a recall rate of 88.84%, and an average precision of 90.04%; the improved YOLOv7 model had a detection accuracy of 87.22%, a recall rate of 90.02%, and an average precision of 96.65%. The harmonic means of the precision and recall rates for the YOLOX, YOLOv5, YOLOv7, and improved YOLOv7 models were 0.84, 0.83, 0.84, and 0.89, respectively. The experimental data were plotted into a PR curve, as shown in Figure 14.

**Table 3.** Research of precision.

| Model | Average Precision (%) | Precision (%) | Recall (%) | F1 Sore |
|---|---|---|---|---|
| YOLOv7-SE-SloU | 95.65 | 87.22 | 90.02 | 0.89 |
| YOLOv7-SloU | 93.86 | 85.98 | 89.73 | 0.88 |
| YOLOv7-SE | 92.54 | 83.63 | 88.85 | 0.87 |
| YOLOv7 | 90.04 | 82.34 | 88.84 | 0.74 |
| YOLOX | 89.65 | 81.91 | 86.35 | 0.64 |
| YOLOv5 | 86.65 | 79.50 | 81.83 | 0.83 |

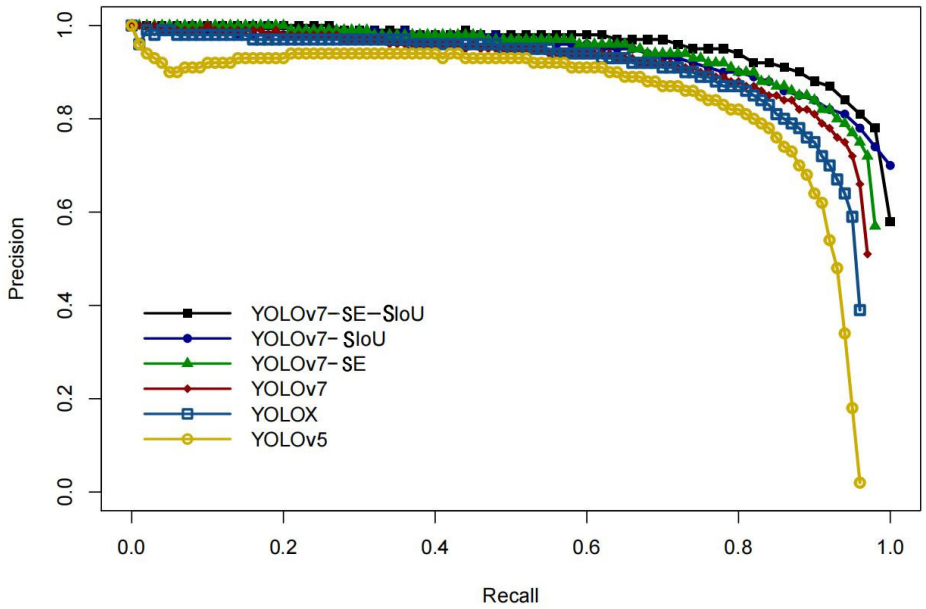


**Figure 14.** PR curves of different detection models.

## 5. Conclusions

The utilization of the improved YOLOv7 model for detecting the coloring and ripening periods of dwarf rootstock apples contributes to effective fruit tree harvest and precise greenhouse management. It promotes the green development of fruit production and strengthens the innovation of smart agricultural technology, thus making a positive contribution to the sustainable development of the apple industry. The proposed model uses an efficient, parameterizable backbone network called RepBlock for lightweight networks, introduces the SE attention mechanism to optimize features, and uses the SIoU loss function to improve the ability to detect obscured apples. The proposed model shows significant improvements in precision, recall, and average accuracy rate, while being lightweight. Compared to the YOLOv7 model, it has achieved remarkable increases of 4.88%, 1.18%, and 5.61% in precision, recall, and average accuracy rate, respectively. In conclusion, the proposed model exhibits significant potential and advantages in detecting the coloring and maturity stages of dwarf rootstock apples. In future research, further optimization of the detection model can enhance its precision and real-time performance. Further research on deploying the improved model into practical applications, such as quantitative fertilization during the coloring stage and real-time harvesting during the ripening periods, can save resources, protect the environment, and improve farming efficiency, thereby promoting the sustainable development of agriculture. Future research and applications will drive advancements in smart agriculture, offering precise management of resources like water, soil, and fertilizers. This reduces biochemical use, lowers labor intensity, and promotes sustainable agricultural development.

**Author Contributions:** Conceptualization, H.M. and X.Z.; methodology, H.M., X.Z. and Y.L. (Yanwen Li); software, H.M. and Y.L. (Yanwen Li); validation, H.M., Y.L. (Yanwen Li) and Y.L. (Yaoyu Li); resources, H.M. and X.Z.; data curation, H.M., Y.L. (Yanwen Li) and X.Z.; writing—original draft preparation, H.M.; writing—review and editing, H.M. and X.Z.; visualization, H.M., Y.L. (Yanwen Li), Z.L., Q.Z., R.Z. and R.H.; supervision, X.Z. and Y.L. (Yanwen Li); funding acquisition, H.M. and X.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research is supported by the Key Research and Development Project in Shanxi Province (No. 202202140601021) and the guiding fund of the award-winning project of "Internet +" Competition of Shanxi Agricultural University (23142N02180001).

**Data Availability Statement:** The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

**Acknowledgments:** We are deeply grateful to Yanwen Li and Jian Wang (Cotton Research Institute, Shanxi Agricultural University) for their valuable support, providing us with an experimental platform and providing us with valuable insights. We would like to thank Zhang Yongbin (Shanxi Fruit tree Research Institute) for his help in providing experimental materials and data sampling. Their participation is essential to ensure the quality and accuracy of our research results.

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

## References

1. Ličina, V.; Krogstad, T.; Fotirić Akšić, M.; Meland, M. Apple Growing in Norway—Ecologic Factors, Current Fertilization Practices and Fruit Quality: A Case Study. *Horticulturae* **2024**, *10*, 233. [CrossRef]
2. Wang, H.; Ba, G.; Uwamungu, J.Y.; Ma, W.; Yang, L. Transcription Factor MdPLT1 Involved Adventitious Root Initiation in Apple Rootstocks. *Horticulturae* **2024**, *10*, 64. [CrossRef]
3. Wang, J.; Xue, L.; Zhang, X.; Hou, Y.; Zheng, K.; Fu, D.; Dong, W. A New Function of MbIAA19 Identified to Modulate Malus Plants Dwarfing Growth. *Plants* **2023**, *12*, 3097. [CrossRef] [PubMed]
4. Li, S.; Zhang, Y.; Chen, H.; Li, B.; Liang, B.; Xu, J. The Effect of Dwarfing Interstocks on Vegetative Growth, Fruit Quality and Ionome Nutrition of 'Fuji' Apple Cultivar 'Tianhong 2'—A One-Year Study. *Plants* **2023**, *12*, 2158. [CrossRef] [PubMed]
5. Wang, H.B.; Zhou, Z.Y.; Yang, Z.F.; Cao, Y.F.; Zhang, C.X.; Cheng, C.G.; Zhou, Z.S.; Wang, W.H.; Hu, C.Z.; Feng, X.J. Analysis of Constrained Factors for High-quality Development of Fruit Industry in China. *China Fruits* **2023**, *7*, 1–9.
6. Vrochidou, E.; Tsakalidou, V.N.; Kalathas, I.; Gkrimpizis, T.; Pachidis, T.; Kaburlasos, V.G. An Overview of End Effectors in Agricultural Robotic Harvesting Systems. *Agriculture* **2022**, *12*, 1240. [CrossRef]
7. Otani, T.; Itoh, A.; Mizukami, H.; Murakami, M.; Yoshida, S.; Terae, K.; Tanaka, T.; Masaya, K.; Aotake, S.; Funabashi, M.; et al. Agricultural Robot under Solar Panels for Sowing, Pruning, and Harvesting in a Synecoculture Environment. *Agriculture* **2022**, *13*, 18. [CrossRef]
8. Jiang, Y. The Impact of Agricultural Technology Innovation on Crop Yield and Agricultural Economic Benefits. *Cotton Sci.* **2023**, *45*, 21–23.
9. Li, Z.; Yang, S.; Shi, D.; Liu, X.; Zheng, Y. A method for determining apple tree yield based on lightweight improved YOLOv5. *Smart Agric.* **2021**, *3*, 1–15. [CrossRef]
10. Jing, P.; Li, M.; Cheng, T.; Sun, N.; Zhang, H.; Xia, X.; Yang, J. Application of Machine Learning in Smart Apple Production. *J. Jilin Agric. Univ.* **2021**, *43*, 138–145.
11. Guo, X.; Hao, Q.; Yang, F. Multi-Target Detection of Apples Based on Improved HOG and SVM. *Foreign Electron. Meas. Technol.* **2022**, *41*, 154–159.
12. Jia, H.; Wang, G.; Guo, M.; Shah, D.; Jiang, X.; Zhao, J.L. Method and experiment of maize plant number acquisition based on machine vision. *Trans. Chin. Soc. Agric. Eng.* **2015**, *31*, 215–220.
13. Kumar, P.S.; Farid, M.; Jonni, M. Deep Learning-Based Apple Detection with Attention Module and Improved Loss Function in YOLO. *Remote Sens.* **2023**, *15*, 1516.
14. Xia, X.; Chai, X.; Li, Z.; Zhang, N.; Sun, T. MTYOLOX: Multi-transformers-enabled YOLO for tree-level apple inflorescences detection and density mapping. *Comput. Electron. Agric.* **2023**, *209*, 107803. [CrossRef]
15. Ma, L.; Zhao, L.; Wang, Z.; Zhang, J.; Chen, G. Detection and Counting of Small Target Apples under Complicated Environments by Using Improved YOLOv7-tiny. *Agronomy* **2023**, *13*, 1419. [CrossRef]
16. Chen, J.; Ma, B.; Ji, C.; Zhang, J.; Feng, Q.; Liu, X.; Li, Y. Apple inflorescence recognition of phenology stage in complex background based on improved YOLOv7. *Comput. Electron. Agric.* **2023**, *211*, 108048. [CrossRef]
17. Zhi, L.; Bao, S.; Kai, B. Optimization of YOLOv7 Based on PConv, SE Attention and Wise-IoU. *Int. J. Comput. Intell. Appl.* **2024**, *23*, 2350033.
18. Zhang, Q.; Liu, Y.; Zhang, Y.; Zong, M.; Zhu, J. Improved YOLOv3 Integrating SENet and Optimized GIoU Loss for Occluded Pedestrian Detection. *Sensors* **2023**, *23*, 9089. [CrossRef] [PubMed]
19. Xia, H.; Tan, L. Research on Improved YOLOv5 Rice Leaf Disease Detection Algorithm, China, Computer and Information Technology. *Comput. Inf. Technol.* **2024**, *32*, 22–25. [CrossRef]
20. Zhang, X.; Zhu, D.; Gan, W. YOLOv7t-CEBC Network for Underwater Litter Detection. *J. Mar. Sci. Eng.* **2024**, *12*, 524. [CrossRef]
21. Yang, T.; Yang, T.; Yang, Y.; Bu, X.; Cao, L. Strawberry target detection method based on an improved YOLOv7. *China Smart Agric. Guide* **2023**, *21*, 005.
22. Liu, Z.; Wei, D.; Li, M.; Zhou, S.; Lu, L.; Dong, X. The orange fruit identification method based on the improved YOLO v5. *China Jiangsu Agric. Sci.* **2023**, *19*, 026.
23. Zhu, J.; Bao, J.; Tao, Y. A Nondestructive Methodology for Determining Chemical Composition of Salvia miltiorrhiza via Hyperspectral Imaging Analysis and Squeeze-and-Excitation Residual Networks. *Sensors* **2023**, *23*, 9345. [CrossRef] [PubMed]

24. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
25. Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; Sun, J. Repvgg: Making vgg-style convnets great again. In Proceedings of the IEEE/CVF Conference on Computer VISION and pattern Recognition, Virtual, 19–25 June 2021; pp. 13733–13742.