electronics

MDPI

Article

Grape Maturity Detection and Visual Pre-Positioning Based on Improved YOLOv4

Chang Qiu [1], Guangzhao Tian [2,*], Jiawei Zhao [2], Qin Liu [3], Shangjie Xie [2] and Kui Zheng [4]

[1] College of Artificial Intelligence, Nanjing Agricultural University, Nanjing 210031, China
[2] College of Engineering, Nanjing Agricultural University, Nanjing 210031, China
[3] School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China
[4] SUNWAY-AI Technology (Changzhou) Co., Ltd., Changzhou 213161, China
* Correspondence: tgz@njau.edu.cn

Abstract: To guide grape picking robots to recognize and classify the grapes with different maturity quickly and accurately in the complex environment of the orchard, and to obtain the spatial position information of the grape clusters, an algorithm of grape maturity detection and visual pre-positioning based on improved YOLOv4 is proposed in this study. The detection algorithm uses Mobilenetv3 as the backbone feature extraction network, uses deep separable convolution instead of ordinary convolution, and uses the h-swish function instead of the swish function to reduce the number of model parameters and improve the detection speed of the model. At the same time, the SENet attention mechanism is added to the model to improve the detection accuracy, and finally the SM-YOLOv4 algorithm based on improved YOLOv4 is constructed. The experimental results of maturity detection showed that the overall average accuracy of the trained SM-YOLOv4 target detection algorithm under the verification set reached 93.52%, and the average detection time was 10.82 ms. Obtaining the spatial position of grape clusters is a grape cluster pre-positioning method based on binocular stereo vision. In the pre-positioning experiment, the maximum error was 32 mm, the mean error was 27 mm, and the mean error ratio was 3.89%. Compared with YOLOv5, YOLOv4-Tiny, Faster_R-CNN, and other target detection algorithms, which have greater advantages in accuracy and speed, have good robustness and real-time performance in the actual orchard complex environment, and can simultaneously meet the requirements of grape fruit maturity recognition accuracy and detection speed, as well as the visual pre-positioning requirements of grape picking robots in the orchard complex environment. It can reliably indicate the growth stage of grapes, so as to complete the picking of grapes at the best time, and it can guide the robot to move to the picking position, which is a prerequisite for the precise picking of grapes in the complex environment of the orchard.

Keywords: machine vision; deep learning; maturation detection; visual pre-positioning; grape; YOLOv4; Mobilenetv3

Citation: Qiu, C.; Tian, G.; Zhao, J.; Liu, Q.; Xie, S.; Zheng, K. Grape Maturity Detection and Visual Pre-Positioning Based on Improved YOLOv4. *Electronics* **2022**, *11*, 2677. https://doi.org/10.3390/electronics11172677

Academic Editor: Pal Varga

Received: 27 July 2022
Accepted: 23 August 2022
Published: 26 August 2022

1. Introduction

China is currently the leading country in the world in terms of grape cultivation production, and the grape industry in China has maintained a rapid growth trend [1]. However, manual harvesting is still used in domestic grape harvesting, which consumes a lot of labor. With the development of computer image processing, machine vision, and control science, it has become possible for automated and intelligent grape picking robots to enter agricultural production [2]. However, the complex environment in grape orchards, the serious shielding of grape branches and leaves to grape fruits, and the overlapping phenomenon between grape fruits make the detection of grape fruit maturity a difficult problem. At the same time, the grape stem also has the problem of being obscured, which makes it difficult to accurately locate the stem position directly when

picking grapes. Therefore, pre-positioning the distribution of grape clusters can guide the robot to move to the picking position, and then accurately locate the picking point, so as to improve the picking efficiency of the robot.

At present, the detection methods of grape maturity have been widely studied at home and abroad. As early as the 1980s, Lee et al. [3] used a puncture method to detect changes in grape hardness throughout the ripening period, but this method of determining grape maturity caused damage to the grape fruit. Julio et al. [4,5] used the near-infrared hyperspectral imaging system to record the hyperspectral images of intact grapes during ripening and used the measured data such as sugar concentration and pH to determine the maturity. However, this method has great limitations and can only be used to detect individual grape grains that have been picked. Behroozi-Khazaei et al. [6] used an artificial neural network (ANN) to achieve the identification of ripe grape clusters, but this method can only be performed under specific light conditions and specific maturity, and the accuracy of the identification needs to be improved. Lu Weiqi [7] used machine vision technology, combined with a C-means clustering algorithm and changes in epidermal color during grape growth, to achieve a more accurate detection of grape maturity, but still can only detect the picked grapes. Therefore, the above method is not suitable for grape maturity detection in complex orchard environments, and it is difficult to apply in practical production.

For the localization of various types of fruit picking, a lot of research has also been conducted at home and abroad. Chen Yan et al. [8] designed the YOLOv3 DenseNet34 litchi string detection network, proposed a litchi string matching method with peer sequence consistency constraints, and calculated the spatial coordinates of litchi strings based on the triangulation principle of binocular stereo vision. Liang Xifeng et al. [9] proposed a calculation method based on the corner of the fruit stalk skeleton and used this algorithm to obtain the position information of the picking point of the fruit stalk of tomato fruit. Mehta et al. [10] used a monocular camera to obtain 3D fruit positions based on a small computationally intensive perspective transformation distance estimation method for real-time robotic control. However, the grape stalk is seriously obscured, which increases the difficulty of direct recognition and positioning of the stalk. Therefore, it is necessary to study the visual pre-positioning of grape picking robots in complex orchard environment.

In recent years, deep convolution neural networks have shown great advantages in target detection, as they can complete the detection task quickly and accurately, making it possible to identify and preposition the fruit maturity in complex environments [11]. In order to solve the grape maturity detection and pre-positioning in complex scenes, the fourth-generation algorithm YOLOv4 of the regression-based YOLO series [12] is selected for grape maturity detection, and the binocular camera is used as the visual sensor for grape cluster pre-positioning. Among them, the improved YOLOv4 algorithm in this study replaces the backbone feature extraction network with Mobilenetv3 network as an improvement, and introduces the SENet attention mechanism, so that it can meet both the accuracy and detection speed requirements of recognition and pre-positioning and compares it with different network models to evaluate the performance and effectiveness of the model.

2. Materials and Methods

The method proposed in this paper consists of two parts: grape maturity detection and grape string pre-positioning, as shown below.

(1) Detection of grape maturity is performed by acquiring images from the left and right cameras of the binocular camera at the same time, and then inputting them into the improved YOLOv4 model for maturity detection.

(2) The pre-positioning of grape clusters is used to match the target through the output detection frame information. After the matching is successful, the parallax

information of the same grape cluster is obtained, and the camera coordinates of grape clusters are obtained using the binocular camera triangulation principle.

2.1. Grape Maturity Detection Based on Improved YOLOv4

2.1.1. Data Collection

The grape images used in this paper were collected in a vineyard in Lianjiang county, Fuzhou City. The grape variety is Kyoho grape. The ear of this grape is generally conical, and the fruit color appears cyan when it is immature, then turns light green, and then gradually turns purple until it is purple–black when it is mature. The image acquisition device is an iPhone 11 (Equipment manufacturer is Apple Inc. of Los Altos, CA, USA.) with a pixel resolution of 4032 × 3024. The pictures were finally saved in JPG format. The images were collected by randomly taking pictures of a grape plant in multiple directions, angles, and distances under different lighting conditions and different overlapping shading levels. Generally, there were multiple grape clusters in each picture, among which there may be grapes with different maturities. Finally, 1000 original images were obtained. According to the maturity of grape clusters, the maturity was defined as four stages: the immature stage of full green, the immature stage of a small number of fruit grains turning purple, the near mature stage of most fruit grains turning purple, and the mature stage of full purple. Figure 1 shows the grape images of four maturity levels.

Figure 1. Images of grapes of four maturity levels.

The environment in the orchard is complex, and the changes in external lighting conditions and the overlapping occlusion of the grapes themselves are diverse. Therefore, this study will discuss the three lighting conditions of natural light, side light, and back light and the three overlapping occlusion conditions of no overlapping occlusion, slight overlapping occlusion, and severe overlapping occlusion. The grape images in complex scenes are shown in Figure 2.

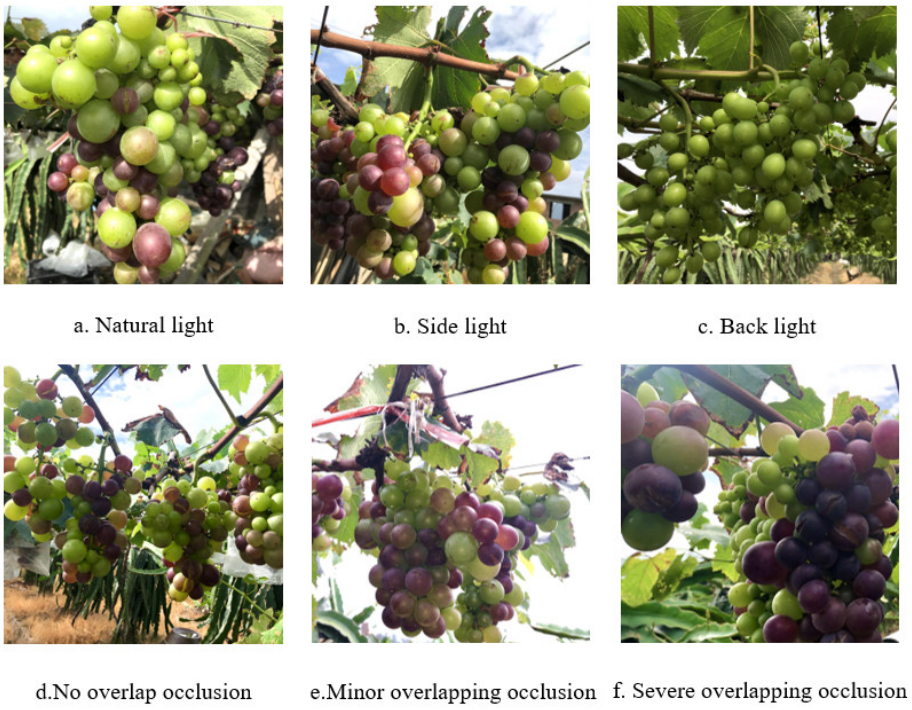

Figure 2. Grape images in complex scenes.

2.1.2. Data Preprocessing

The training model in this study used the dataset format of PASCAL VOC and labeled the targets using LabelImg software. (Software version number is LabelImg 1.8.6, created by tzutalin. The creator has released the version of the software in Canada). Among them, the immature stage of full green was marked as young_ grape (red box), the immature stage of a small number of fruit grains turning purple was marked as near_ young_ grape (green box), the near mature stage of most fruit grains turning purple was marked as near_ mature_ grape (blue box), and the mature stage of full purple was marked as nature_ grape (purple box).

The images of unripe grapes containing full green and ripe grapes with full purple in the dataset were greater and relatively similar in number, but the images of the other two ripening stages were less. Therefore, in order to solve the problem of low recognition rate caused by the large difference in the number of samples of grapes with different maturity [13], a total of 200 grape images with corresponding maturity were found on the internet, so that the number of samples of grapes with four maturity levels can be closer and saved in jpg format uniformly.

To prevent overfitting of the network model due to the small amount of data in the dataset [14], and to enhance the model training effect and model generalization ability [15], this study used the 1200 grape images obtained above to randomly perform mirror inversion, noise addition, translation, and other operations, so as to expand them to 2000, of which there were 7390 grape clusters in total. After completing the data expansion, they were randomly divided into a training set (6651 images) and a validation set (739 images) according to the ratio of 9:1.

2.1.3. Network Structure of the YOLOv4 Algorithm

In 2016, Redmon J et al. proposed a single-stage target detection algorithm based on deep learning, which discards candidate box extraction and uses the regression method to directly classify and predict targets [16]. YOLOv4 is the fourth-generation algorithm of YOLO series. Although it can ensure high recognition accuracy, its computing speed needs to be strengthened.

The backbone feature extraction network of YOLOv4 is CSPDarknet53, which extracts image features by convolving the input image several times and finally obtains three effective feature layers of sizes 52 × 52 × 256, 26 × 26 × 512, and 13 × 13 × 1024, respectively. The last effective feature layer of size 13 × 13 × 1024 is transmitted into the enhanced feature extraction network SPP after three convolutions. The SPP network pools the incoming feature layer by using the maximum pool of four different pool core sizes, which can greatly increase the receptive field, separate the most significant context features, and then transmit them to another feature extraction network PAN after stacking and three convolutions. The feature layers transmitted into the PAN network are first stacked with the effective feature layers of corresponding sizes after two upsampling procedures and then two downsampling procedures, using repeated feature extraction to obtain better features, and finally transmitted to YOLO Head for prediction [17].

2.1.4. Characteristics of the YOLOv4 Algorithm

YOLOv4 uses the mosaic data enhancement method, as shown in Figure 3. This method stitches four pictures into one picture after flipping, scaling, gamut change, and other operations, which can enrich the background of the detection target. In addition, the data of four images will be calculated at the same time during the standardized BN (batch normalization) calculation, which is equivalent to increasing the Batchsize, making the mean and variance calculated by the BN layer closer to the distribution of the overall dataset. Therefore, this method can enhance the robustness and generalization ability of the network [18].

Figure 3. Mosaic data enhancement.

In order to balance the training errors of the prediction frame, confidence, and category, YOLOv4 uses the complete intersection union ratio CIoU as the loss function [19]. Ordinary IoU cannot directly optimize the part without an overlap, while CIoU can more accurately reflect the distance and coincidence between the target frame and the prediction frame and take the ratio of the length and width of the prediction frame as a penalty item, which makes the regression of the target frame more stable, and solves the problems of divergence in the training process of IoU and GIoU [20]. The calculation method of CIoU is shown in Equation (1).

$$\mathrm{CIoU} = \mathrm{IoU} - \frac{\rho^2(b, b^{gt})}{c^2} - \alpha v \quad (1)$$

Among them, ρ^2 (b, b^{gt}) represents the Euclidean distance between the center points of the prediction box and the real box, and c represents the diagonal distance of the smallest closure region that can contain both the prediction box and the real box. α is a

trade-off parameter, v refers to the aspect ratio, and their calculation methods are shown in Equations (2) and (3):

$$\alpha = \frac{v}{1 - \mathrm{IoU} + \mathrm{v}} \tag{2}$$

$$v = \frac{4}{\pi^2}\left(arctan\frac{\omega^{gt}}{h^{gt}} - arctan\frac{\omega}{h}\right)^2 \tag{3}$$

Finally, the loss function is obtained as shown in Equation (4):

$$\mathrm{Loss}_{\mathrm{CIoU}} = 1 - \mathrm{IoU} + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \tag{4}$$

The CIoU loss value comprehensively considers the overlapping area, the distance between the center points, and the length width ratio, which can make the regression of the target box more stable, the convergence speed faster, and the position prediction more accurate. The model can be optimized by back propagation even if the predicted frame does not intersect with the real frame.

YOLOv4 uses the cosine annealing attenuation method to reduce the learning rate, so that the network is as close to the global minimum of the loss value as possible, and it can converge to the optimal solution. Cosine annealing, as the name implies, is used to reduce the learning rate through the cosine function, and the gradient descent value changes according to the change in the cosine function value. Using this method, the network can find the global optimal solution instead of the local optimal solution.

The principle of the cosine annealing attenuation method is shown in Equation (5).

$$\eta_t = \eta_{min}^i + \frac{1}{2}(\eta_{max}^i - \eta_{min}^i)(1 + cos(\frac{T_{cur}}{T_i}\pi)) \tag{5}$$

Among them, η_t is the current learning rate; i is the number of index values; η_{max}^i and η_{min}^i represent the maximum and minimum learning rates, respectively; T_{cur} is the number of cycles currently executed; and T_i is the total number of cycles in the current operating environment.

2.1.5. Improvement of the YOLOv4 Algorithm

The improvement of YOLOv4 in this study is that Mobilenetv3 is used as the backbone feature extraction network to replace the original CSPDarknet53, and the idea of implementing the backbone feature extraction network replacement is to replace the three preliminary effective feature layers with the same size. In order to further reduce the number of parameters, the depth separable convolution is used to replace the ordinary convolution in the enhanced feature extraction network. At the same time, the SENet attention mechanism is added before some features are fused.

The Mobilenet model is a lightweight deep neural network proposed by Google. It uses a special Bneck structure, which combines the deep separable convolution of Mobilenetv1 and the linear bottleneck inverse residual structure of Mobilenetv2. At the same time, the lightweight attention module is introduced into some Bneck structures, which increases the channel weight with strong feature extraction ability. In the backbone module, the h-swish activation function is used to replace the swish function, which reduces the amount of network calculation [21,22].

The specific implementation process of Mobilenetv3 is shown in Figure 4. First, the input feature layer is up-dimensioned using 1 × 1 convolution to expand the number of channels of the input feature layer, and then feature extraction is performed using 3 × 3 depth-separable convolution, and then the attention mechanism is added to it by one global average pooling, two full connections, and one multiplication operation, and finally after being down-dimensioned using 1 × 1 convolution, the construction of the backbone

part is completed. The construction of the residual edge of the model only needs to connect the input and output directly.

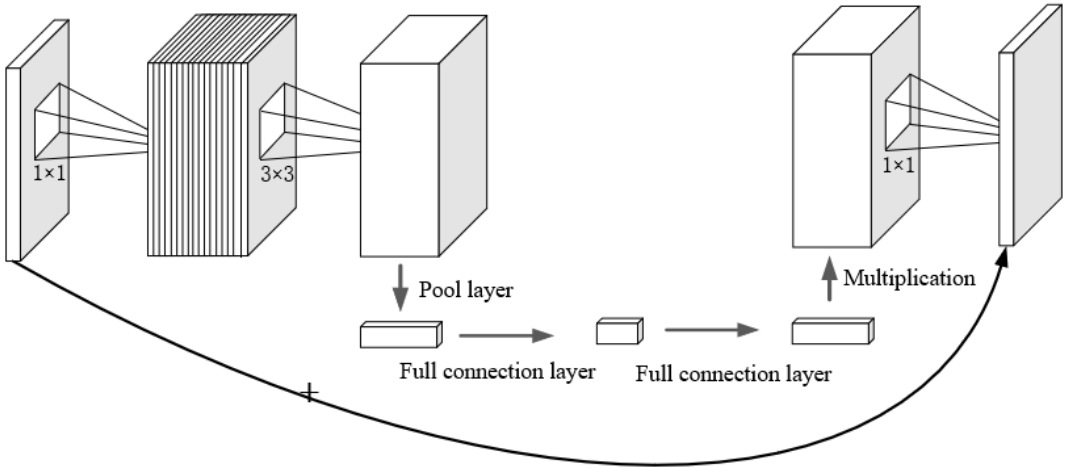

Figure 4. Bneck structure schematic diagram in MobilenetV3.

The attention mechanism is a way to realize network adaptive attention, and its core focus is to make the network pay attention to the places that need more attention. Attention mechanisms are generally divided into channel attention mechanisms and spatial attention mechanisms. This study uses the channel attention mechanism SENet, which leads the network to pay attention to the channel that needs attention most. The specific implementation method is shown in Figure 5 below. First, the input feature layer is global average pooling, and then two full connections are made. The number of neurons in the first full connection is small, and the number of neurons in the second full connection is the same as that in the input feature layer. After completing two full connections, the sigmoid is taken again to fix the value between 0 and 1. At this time, the weight of each channel of the input feature layer can be obtained, and finally the weights are multiplied by the original input feature layer.

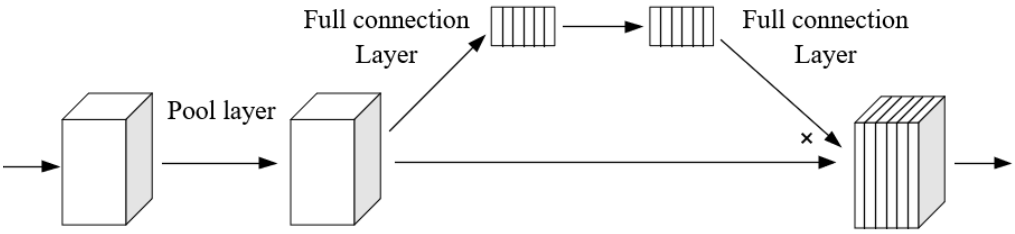

Figure 5. Implementation of the SENet attention mechanism.

The improved YOLOv4 network model is shown in Figure 6.

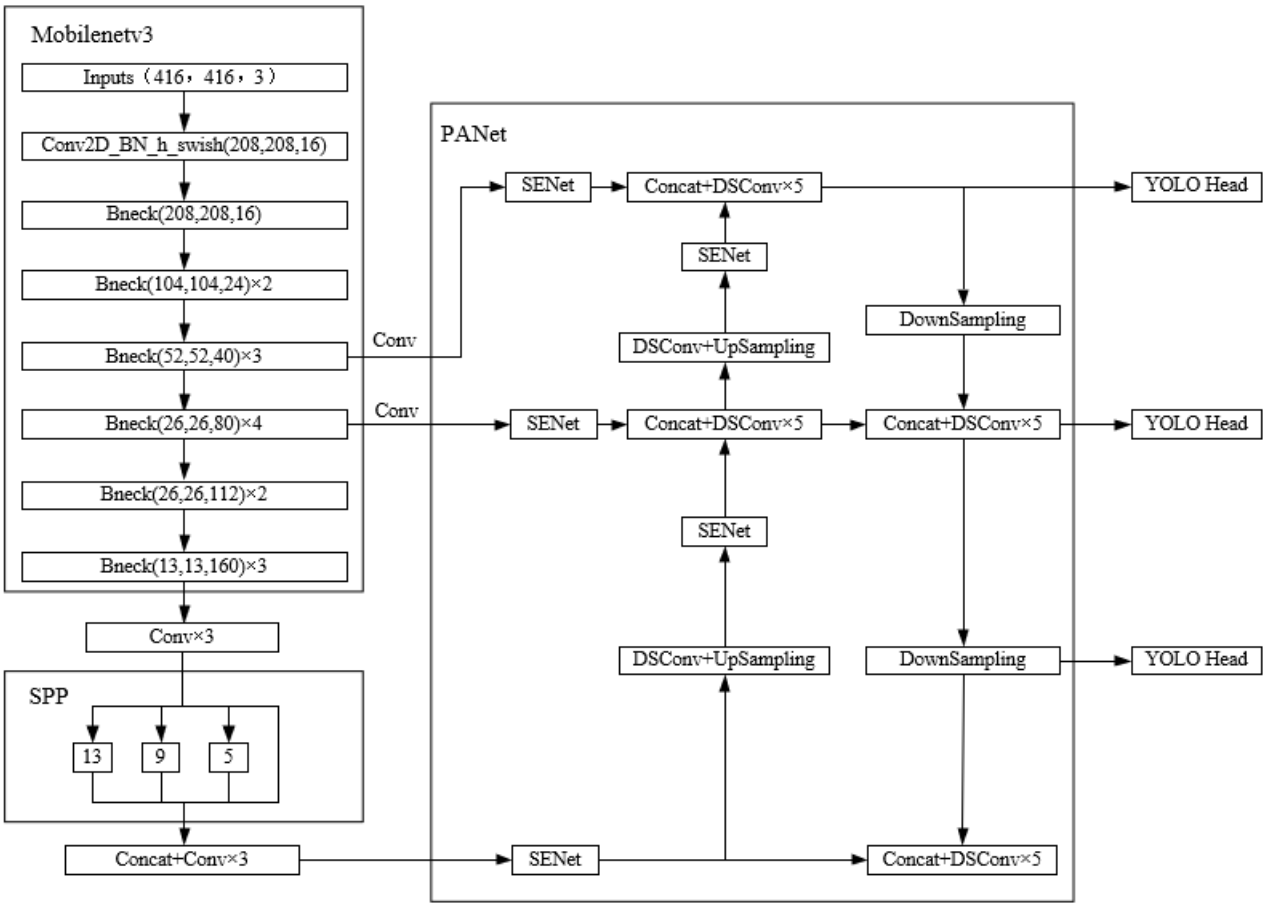

Figure 6. Framework of the improved YoloV4 network model. Conv2D_BN_h_swish represents a composite module with 2D convolution (Conv2D) plus bulk regular (BN) plus h_swish activation function; Conv represents convolution; DSConv represents deep separable convolution; Concat represents stack; UpSampling represents upsampling; DownSampling represents downsampling.

2.2. Grape Cluster Pre-Positioning Based on Binocular Stereo Vision

2.2.1. Target Matching

Firstly, the maturity of the left and right images collected by the binocular camera is detected. Then, the corresponding points of the two images are stereo matched. The specific process is as follows [23].

(1) Obtain the target detection frames ($BBOX_L$) and ($BBOX_R$) of the left and right images (PIC_L) and (PIC_R) and their categories ($CLASS_L$) and ($CLASS_R$).

(2) Calculate the pixel areas (S_L) and (S_R) through the width (w) and height (h) of the target detection frame, and then calculate the pixel coordinates (C_L) and (C_R) of the centroid of the target detection frame, so as to obtain the difference between the pixel area (S_D) and the centroid ordinate (V_D).

(3) Judge whether ($CLASS_L$) and ($CLASS_R$) are the same and whether (S_D) and (V_D) are less than the threshold value.

If it is satisfied at the same time, the matching is successful, otherwise, it fails. Among them, according to experiments and experiences, the thresholds value of the difference between the pixel area and the centroid ordinate are set to 0.02 (S_L) and 0.01 (V_L), respectively [24].

If the detection frames of the left and right images are successfully matched, it means that the detection of grape maturity in the two images is consistent and the spatial location of grapes is obtained. If the matching fails, the above image capture and detection work should be repeated until the matching is successful.

2.2.2. Pixel Parallax Calculation

The centroid of the successfully matched target detection frame is used to represent the grape cluster, as shown in Figure 7. The centroid coordinates of the left and right detection frames are set to (U_L, V_L) and (U_R, V_R), respectively. Since the left and right eyes of the binocular camera are on the same horizontal line, the difference between the two centroid coordinates is the difference between the horizontal coordinates, and the pixel parallax (D) of the same grape cluster in the left and right destination pixels is as shown in Equation (6).

$$D = U_L - U_R \tag{6}$$

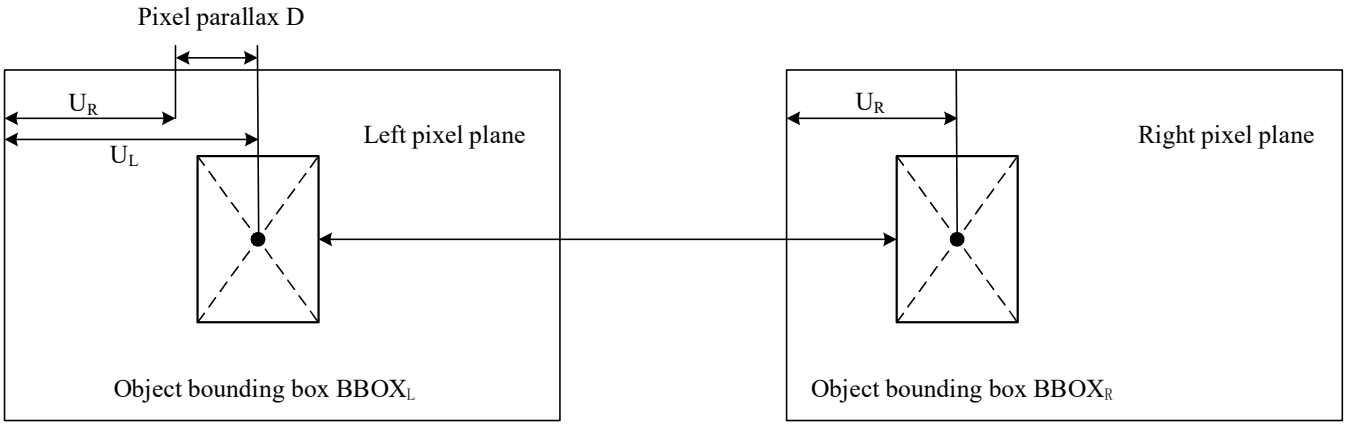

Figure 7. Pixel parallax.

2.2.3. Depth Estimation of Grape Clusters

The left and right camera aperture centers of binocular cameras are noted as O_L and O_R, the focal length of the camera is noted as f, and the baselines of the camera are noted as b. Horizontal coordinates of the target detection frame in the image coordinate system are noted as X_L and X_R, and the difference in horizontal coordinates is the parallax d. Then, according to the similarity principle of the triangle, the depth (Z) can be obtained as:

$$Z = \frac{fb}{d} = \frac{fb}{X_L - X_R} \tag{7}$$

The horizontal translation of the origin of the image coordinate system in the pixel coordinate system is noted as U_0, and the scale factor of the U-axis is noted as f_x. Then, according to the conversion principle between the pixel coordinates and the image coordinates, the horizontal coordinates (U_L) and (U_R) in the pixel coordinate system can be obtained as follows:

$$\begin{cases} U_L = \frac{f_x}{f} X_L + U_0 \\ U_R = \frac{f_x}{f} X_R + U_0 \end{cases} \tag{8}$$

The depth (Z) obtained by simultaneous Equations (7) and (8) is:

$$Z = \frac{f_x \cdot b}{U_L - U_R} \tag{9}$$

If the coordinates of the grape cluster are (X, Y, Z) for the binocular camera, it can be known from the depth (Z),

$$\begin{cases} X = \dfrac{Z\ (U_L - U_0)}{f_x} \\ Y = \dfrac{Z\ (V_L - V_0)}{f_y} \end{cases} \tag{10}$$

3. Grape Maturity Detection and Pre-Positioning Test

3.1. Test Platform and Evaluation Index of Grape Maturity Detection

The specific configuration of the deep learning environment in this study was Intel Corei5 12400F 4.40 GHz CPU; 16 G running memory; 500 G solid state disk; 11 GB NVIDIA GeForce GTX 1080 Ti GPU; CUDA and Cudnn versions were 10.0 and 7.6.5, respectively.

The parameters of the network training were set as follows. The number of samples of iterative training was 8, the total number of iterations was 2000, the initial learning rate was 0.001, the momentum factor was 0.95, and every 50 training sessions a training weight was saved and the learning rate was reduced by a factor of 10.

In this study, Precision–Recall curve, AP value (detection accuracy), MAP (mean value of AP value under all classes, which is divided into four classes in this study), and detection speed were used as evaluation indexes. Among them, AP value is the area enclosed by the P-R curve and coordinate axis, and MAP is the average value of AP of all classes [25].

After the network training was completed, the prediction frame finally selected by the network may appear in the following three situations. The first situation was that the prediction frame hits the real target frame, and the number of such cases was represented by TP. The second case was that the prediction frame does not hit the real target frame, and the number of such cases was represented by FP. The third case was that there is no prediction frame in the real target area, and the number of such cases was represented by FN [26]. Among them, P and R can be expressed by TP, FP, and FN as shown in Equations (11) and (12), while AP and MAP are related to P and R as shown in Equations (13) and (14).

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \times 100\% \tag{11}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100\% \tag{12}$$

$$\mathrm{AP} = \int_0^1 P(R)\, dR \times 100\% \tag{13}$$

$$\mathrm{MAP} = \frac{\sum_{n=1}^{4} \mathrm{AP}(n)}{4} \times 100\% \tag{14}$$

3.2. Parameter Calibration and Evaluation Index of the Grape Pre-Positioning Test

This experiment used a ZED2i binocular camera with a focal length (f) of 1.8 mm and a base distance (b) of 120 mm, and the resolution of left and right images collected by the camera was 2560 × 720. The binocular camera was calibrated in pixels using the ZED camera's own software development kit SDK. The results are shown in Table 1.

Table 1. Calibration results of the binocular camera.

Cameras	U-Axis Scale Factor	V-Axis Scale Factor	U-Axis Translation	V-Axis Translation
Left camera	529.88	529.54	645.25	367.12
Right camera	529.62	529.45	645.87	384.52

The pre-positioning test of grape clusters was carried out in an orchard in Lianjiang county, Fuzhou City, where several steel wires were installed 2 m above the ground, parallel to each other and forming a plane parallel to the ground. In the early stage of grape growth, the grape branches were installed on the steel wires, so that the fruit grows naturally downward until maturity, and the centroid of each bunch of grapes can be as close to a horizontal plane as possible, which is conducive to the progress of the grape depth estimation experiment.

During the experiment, the left and right optical centers of the binocular camera were kept at the same level as the center of mass of the grape bunches as much as possible, and the connecting line of the left and right optical centers of the binocular camera was parallel to the ground. In the experiment, grape cluster images at 10 locations were randomly collected for maturity detection and depth estimation, and these 10 images were input into the SM-YOLOv4 model and the other two models with better accuracy in subsequent maturity detection.

The distance between the trunk of the fruit tree and the origin of the camera coordinate system measured using a laser rangefinder is (D_{di}). The camera coordinates (X_i, Y_i, Z_i) of the grape clusters are obtained through target matching and parallax information calculation, where X_i and Z_i are the estimated values of horizontal and vertical distances, respectively, and the estimated value of the distance between the grape clusters and the camera (D_i) is:

$$D_i = \sqrt{X_i^2 + Z_i^2} \tag{15}$$

In the experiment, the mean error value (E_D) and the mean error ratio (E_{DR}) are used as the evaluation indicators of positioning accuracy, and the calculation equations are as follows:

$$E_D = \frac{\sum_{i=1}^{n} |D_{di} - D_i|}{n} \tag{16}$$

$$E_{DR} = \frac{\sum_{i=1}^{n} \left| \frac{D_{di} - D_i}{D_{di}} \right|}{n} \times 100\% \tag{17}$$

where n is the number of successfully matched grape clusters in the collected image.

4. Results and Analysis

4.1. Training Results of the Grape Maturity Detection Model

The loss curve and Precision–Recall curve of the model after training are shown in Figure 8. The number of iterations was set to 2000, and the loss value of the model gradually tended to be stable at the 100th iteration, indicating that the network had been fitted and trained well at this time. When the IoU threshold was set to 0.5, the MAP value of the model reached 93.52%, indicating that the detection accuracy of the network can meet the needs.

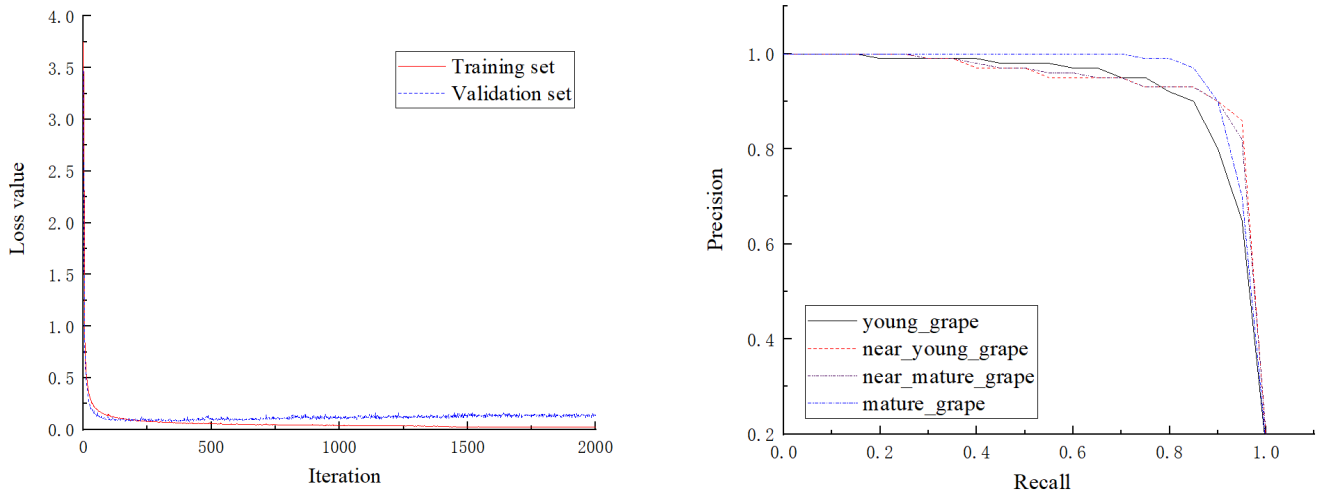

Figure 8. Loss curves and the Precision–Recall curves. (**a**) Loss curve. (**b**) Precision–Recall curve.

4.2. Maturity Test Results of the SM-YOLOv4 Network

4.2.1. Comparison of Training Results of the Improved Network Model

In this paper, we proposed the SM-YOLOv4 algorithm, and the main improvement lay in replacing the backbone feature extraction network of the original YOLOv4 algorithm with Mobilenetv3 and adding the SENet attention mechanism. In order to verify the feasibility of the improved algorithm, ablation experiments were designed to compare the results. Among them, the SE-YOLOv4 algorithm only added the SENet attention mechanism to the YOLOv4 algorithm. The Mobilenetv3-YOLOv4 algorithm was used to replace the backbone feature extraction network of the YOLOv4 algorithm with the lightweight backbone feature extraction network Mobilenetv3. The results of different network models after training are shown in Table 2. According to the results, the MAP of the SM-YOLOv4 algorithm is 93.52%, which is 3.93, 0.89, and 1.58 percentage points higher than the original YOLOv4 algorithm, SE-YOLOv4 algorithm, and Mobilenetv3-YOLOv4 algorithm, respectively. At the same time, the detection rate of the SM-YOLOv4 algorithm can also be guaranteed. Therefore, it can be seen that adding the attention mechanism can improve the average accuracy of the model, using Mobilenetv3 as the backbone feature extraction network can improve the detection speed of the model, and replacing the backbone feature extraction network and adding the SENet attention mechanism simultaneously can achieve better results for all training results of the network, which also shows that the improvement of the network model in this study can meet the needs of both accuracy and speed.

Table 2. Comparison of the training results of the modified network model.

Models	AP/%				MAP/%	Speed /(Frames·s^{-1})
	A	B	C	D		
YOLOv4	84.33	90.76	91.10	92.18	89.59	50.09
SE-YOLOv4	90.32	92.05	93.80	94.36	92.63	47.63
Mobilenetv3-YOLOv4	89.28	92.56	92.43	93.49	91.94	97.87
SM-YOLOv4	90.62	93.95	94.29	95.24	93.52	92.38

Note: AP indicates the average accuracy of grape maturity detection, where A indicates the immature stage of full green, B indicates the immature stage of a small number of fruit grains turning purple, C indicates the near mature stage of most fruit grains turning purple, and D indicates the mature stage of full purple.

4.2.2. Comparison of Training Results of Different Network Models

In this paper, the proposed SM-YOLOv4 model is compared with YOLOv5, YOLOv4-Tiny, and Faster_ R-CNN models. Among them, the network structure of SM-YOLOv4 model was specifically presented in the above sections and will not be repeated here. The other three models will be briefly introduced here.

The YOLOv5 target detection algorithm is the fifth-generation algorithm of the YOLO series. Its core idea is to take the whole picture as the input of the network, and directly use the regression method to obtain the location coordinates and categories of targets at the output layer. Its characteristics are high detection accuracy and fast detection speed, which meet the needs of real-time monitoring.

The YOLOv4-Tiny target detection algorithm is simplified on the basis of YOLOv4, omitting the SPP network and PAN network, thus, reducing the network calculation amount and greatly improving the detection speed of the model. However, the disadvantage of this network is that its detection accuracy needs to be improved.

Faster_R-CNN target detection algorithm is a two-stage target detection algorithm based on region recommendation, which can be regarded as a combination of the region generation network (RPN) and Fast_R-CNN, where the RPN network generates the candidate area and Fast_R-CNN is used for target detection. Compared with the YOLO series algorithms mentioned above, the two-stage algorithms are more complex and have higher detection accuracy, but the detection speed and training speed need to be improved.

The experiment results of the SM-YOLOv4 model and other models on the validation set for different maturity grape detection are shown in Table 3. It can be seen from Table 3 that the overall average accuracy of the SM-YOLOv4 model in detection is higher than that of YOLOv5, YOLOv4-Tiny, and Faster_R-CNN increased by 0.93, 11.06, and 0.41 percentage points, respectively, and the detection rate is also high. At the same time, it can be found that all models have the lowest average accuracy for the detection of full green grapes, probably because the color of grapes of this maturity is most similar to the color of its surrounding leaves [27], which led to the phenomenon that the model may not be able to distinguish the grape fruit from the leaves during the detection and missed the detection. However, this does not affect the picking of mature fruits. Therefore, the comprehensive comparison shows that the improved model in this paper has greater advantages in the accuracy and speed of detecting grape maturity.

Table 3. Results of grape maturity detection for different trained network models.

Models	AP/%				MAP/%	Speed /(Frames·s^{-1})
	A	B	C	D		
YOLOv5	88.53	93.96	93.30	95.38	92.79	34.79
SM-YOLOv4	90.62	93.95	94.29	95.24	93.52	92.38
YOLOv4-Tiny	75.19	80.80	86.26	87.60	82.46	288.01
Faster_ R-CNN	91.20	92.90	93.89	94.43	93.11	14.53

Note: AP indicates the average accuracy of grape maturity detection, where A indicates the immature stage of full green, B indicates the immature stage of a small number of fruit grains turning purple, C indicates the near mature stage of most fruit grains turning purple, and D indicates the mature stage of full purple.

In order to verify the accuracy, rapidity, and versatility of this model for grape maturity detection in a complex orchard environment, six grape images from various complex scenes found by the network were selected and their maturities were detected. Figure 9 shows the experiment results of grape maturity detection by different models under different conditions. It can be intuitively found from the figure that when the illumination condition was backlight and the fruit seriously overlapped, the accuracy of detection all models was affected, because the texture of the grape surface will become

unclear with backlighting [28], which increased the difficulty of grape detection. In addition, unlike other fruits, bunches of grapes are also more difficult for the human eye to distinguish grapes when overlapping occurs. Therefore, in the case of serious overlapping, the average accuracy of model detection will be greatly affected. According to Table 3 and Figure 9, the YOLOv4-Tiny model has the fastest detection speed but the worst detection accuracy and is prone to missed detection. Faster_R-CNN has high detection accuracy, but it is prone to false detection and the detection speed is the slowest. Therefore, compared with other models, the SM-YOLOv4 model can greatly avoid the missed detection and false detection of other models [29].

Figure 9. Detection of grape maturity with different models under different conditions. Among them, the correct color of detection frame should comply with the following regulations, the immature stage of full green was marked as young_ grape (red box), the immature stage of a small number of fruit grains turning purple was marked as near_ young_ grape (green box), the near mature stage of most fruit grains turning purple was marked as near_ mature_ grape (blue box), and the mature stage of full purple was marked as nature_ grape (purple box).

In order to further illustrate the practicability of the algorithm in this paper, the results of the algorithm proposed in this paper are compared with those of the maturity detection methods proposed in [30,31]. The accuracy of the method in [30] for citrus maturity detection is 95.07%, and the detection time is 23 ms. The accuracy of the method in [31] for coconut maturity detection is 89.4%, and the detection time is 3.124 s. The overall average accuracy of this algorithm is 93.52%, and the average detection time is

10.82 ms. Through comparison, it can be shown that the detection accuracy of this algorithm basically meets the needs of fruit maturity detection and has a high detection speed. It can meet the needs of a grape picking robot in terms of maturity detection accuracy and real-time monitoring.

4.3. Depth Estimation Results of the SM-YOLOv4 Network

The SM-YOLOv4, YOLOv5, and Faster_R-CNN models with high maturity detection accuracy were used to detect and pre-position the grape images collected at 10 random positions. Since the grape picking robots need to pick mature grapes, the pre-positioning experiments were carried out during the mature period of grapes. In addition, the distance between rows of grapes is about 2 m, and the robot walks between two rows of grapes, so the actual distance between the binocular camera and the grape clusters photographed in the pre-positioning experiment ranges from 0 to 2 m.

There were 37 clusters of grapes in the images collected by the binocular camera at 10 locations. Among them, the SM-YOLOv4 model had no false detection and missed detection in target detection, the YOLOv5 model missed three bunches but no false detection occurred, and the Faster_R-CNN model had no missed detection, but one false detection occurred.

As shown in Figure 10, the comparison of the positioning results of different models on grape clusters at positions 1, 3, and position 10 are shown. There are five clusters of grapes in position 1, and the actual distances from left to right are 0.952, 0.839, 0.703, 0.613, and 0.529 m. There are three clusters of grapes in position 3, and the actual distances from left to right are 0.534, 0.375, and 0.425 m. There are three clusters of grapes in position 10, and the actual distances from left to right are 0.556, 0.517, and 0.622 m.

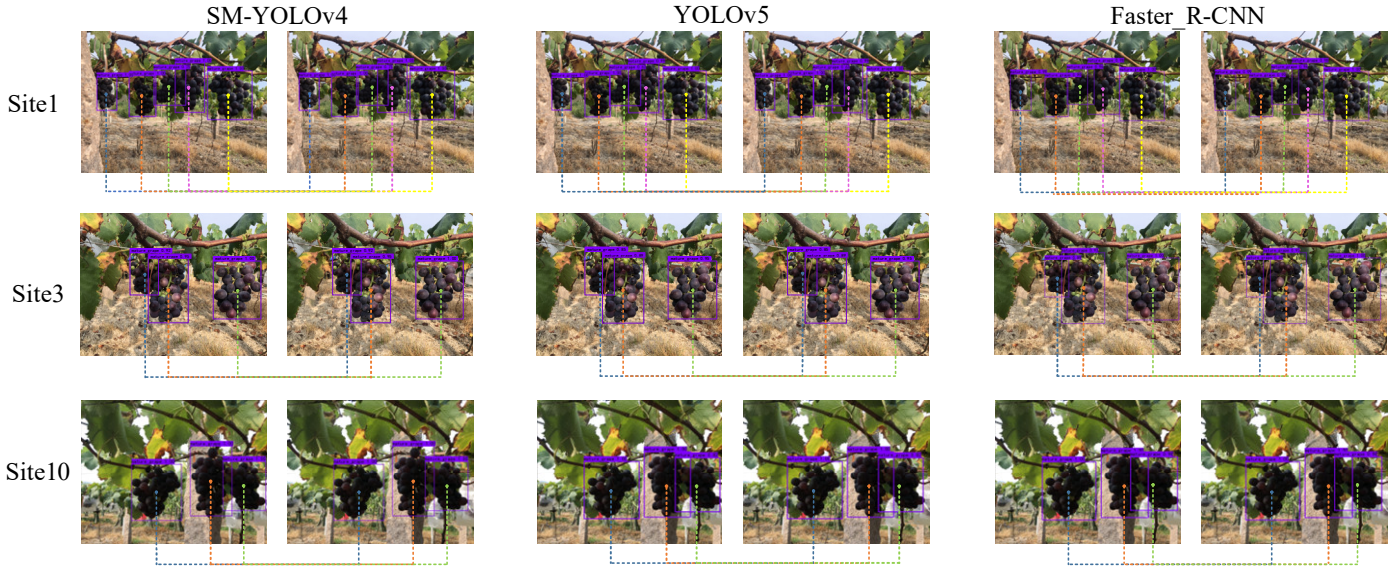

Figure 10. Comparison of the localization results of grape strings using different models.

As shown in Table 4, the results of pre-positioning accuracy of different models for grape clusters are presented. It can be seen from the table that the effect of the pre-positioning experiment using the SM-YOLOv4 model is the best, and the mean error and mean error ratio relative to the YOLOv5 and Faster_R-CNN models have been reduced. The mean error and the mean error ratio of the SM-YOLOv4 model are 0.027 m and 3.89%, respectively, which are 0.04 m and 5.35 percentage points lower than those of the YOLOv5 model, and 0.026 m and 3.49 percentage points lower than those of the Faster_R-CNN model, respectively. Since the detection speed of the SM-YOLOv4 model is also the fastest of the three models, it shows that the application of the SM-YOLOv4 model can effectively pre-position grapes.

Table 4. Grape bunch pre-positioning results.

Site	SM-YOLOv4		YOLOv5		Faster_R-CNN	
	E_D/m	E_{DR}/%	E_D/m	E_{DR}/%	E_D/m	E_{DR}/%
1	0.032	4.63	0.069	9.55	0.057	8.15
2	0.023	3.31	0.061	8.01	0.046	6.19
3	0.032	4.60	0.074	11.00	0.062	9.04
4	0.024	3.45	0.068	9.52	0.052	7.26
5	0.029	4.17	0.066	9.06	0.053	7.43
6	0.030	4.31	0.065	8.83	0.047	6.37
7	0.025	3.59	0.063	8.37	0.049	6.72
8	0.023	3.38	0.062	8.14	0.055	7.79
9	0.021	3.02	0.067	9.29	0.050	6.90
10	0.031	4.46	0.073	10.67	0.056	7.97
Average value	0.027	3.89	0.067	9.24	0.053	7.38

It can be seen from the above results that the detection accuracy of the model largely determines the accuracy of the detection frame, that is, the accuracy of the grape centroid location [32,33], and the determination of the grape centroid location further affects the accuracy of the pre-positioning [34]. Therefore, the SM-YOLOv4 model is outstanding in terms of grape recognition and detection, making its accuracy in the grape pre-positioning test also the most consistent with the actual needs of all models mentioned in the paper.

In order to further illustrate the practicability of the algorithm in this paper, the algorithm proposed in this paper is compared with the pre-positioning method proposed in the literature [35]. In [35], the YOLOv3-DenseNet34 model is used to preposition the litchi string, so as to guide the picking task of the picking robot. The mean error of this method for the pre-positioning of litchi is 0.023 m. The mean error of the method used in this paper is 0.027 m, which shows that the algorithm can meet the precision requirements of the pre-positioning and can be applied in the actual orchard environment.

5. Conclusions

This paper proposed an improved target detection algorithm of SM-YOLOv4, which is used to detect grape maturity in complex orchard environments. In the YOLOv4 network, the backbone feature extraction network is replaced by Mobilenetv3, and the SENet attention mechanism is integrated into the enhanced feature extraction network, which not only enhances the robustness of the network model, but also makes the network more lightweight.

(1) The overall average accuracy of this method under the verification set reaches 93.52%, and the average detection speed is 10.82 ms. Compared with the models before and after the improvement, the overall average accuracy of the improved model is 4.93 percentage points higher than the original model, and the average detection rate is 42.29 ms higher, indicating that the improved model has greater advantages in both accuracy and speed to meet the detection of grape maturity in complex environments.

(2) In order to verify the detection effect of this model in the actual orchard complex environment, different models were used to detect the maturity of grapes under different conditions. The experimental results show that the SM-YOLOv4 model provides higher detection accuracy and faster detection speed compared with YOLOv5, YOLOv4-Tiny, and Faster_R-CNN models under various lighting conditions and overlapping occlusion.

(3) In the pre-positioning experiment, the mean error and the mean error ratio of this model are 0.027 m and 3.89%, respectively, which are lower than YOLOv5 and Faster_R-CNN models, which can effectively guide the grape picking robot for grape pre-positioning.

Author Contributions: Conceptualization, C.Q., G.T.; methodology, C.Q., G.T.; data curation, C.Q., J.Z. and Q.L.; validation, C.Q., K.Z., S.X.; writing—original draft preparation, C.Q., G.T.; writing—review and editing, C.Q., J.Z. and Q.L.; funding acquisition, G.T.; visualization, C.Q., G.T.; supervision, C.Q., G.T.; project administration, C.Q., G.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The National Natural Science Foundation of China (31401291) and The 10th Batch of Changzhou Science and Technology Planning Projects (International Science and Technology cooperation/Hong Kong, Macao and Taiwan Science and Technology cooperation) (CZ20220010).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Luo, S. Grape Industry and the Tourism Development in China. Master's Thesis, Hubei University, Wuhan, China, 2017.
2. Xu, R.; Zhao, M.; Zhang, L. Research actuality and prospect of picking robot for grapes. *J. Henan Inst. Sci. Technol.* **2018**, *46*, 74–78.
3. Lee, C.Y.; Bourne, M.C. Changes in grape firmness during maturation. *J. Texture Stud.* **1980**, *11*, 163–172.
4. Nogales-Bueno, J.; Hernández-Hierro, J.M.; Rodríguez-Pulido, F.J.; Heredia, F.J. Determination of technological maturity of grapes and total phenolic compounds of grape skins in red and white cultivars during ripening by near infrared hyperspectral image: A preliminary approach. *Food Chem.* **2014**, *152*, 586–591.
5. Yuan, L. Study on Non-destructive Detection of 'Kyoho' Grape's Quality by Multi-perspective Imaging and Nir Spectroscopy Techniques. Ph.D. Dissertation, Jiangsu University, Zhengjiang, China, 2016.
6. Behroozi-Khazaei, N.; Maleki, M.R. A robust algorithm based on color features for grape cluster segmentation. *Comput. Electron. Agric.* **2017**, *142*, 41–49.
7. Lu, W. *Grape Maturity's Nondestructive Detection Research*; China Metrology Institute: Beijing, China, 2013.
8. Chen, Y.; Wang, J.; Zeng, Z. Vision pre-positioning method for litchi picking robot under large field of view. *Trans. Chin. Soc. Agric. Eng.* **2019**, *35*, 7.
9. Liang, X.; Jin, C.; Ni, M.; Wang, Y. Acquisition and experiment on location information of picking point of tomato fruit clusters. *Trans. Chin. Soc. Agric. Eng.* **2018**, *34*, 7.
10. Mehta, S.S.; Burks, T.F. Vision-based control of robotic manipulator for citrus harvesting. *Comput. Electron. Agric.* **2014**, *102*, 146–158.
11. Zhu, Y.; Zhou, W.; Yang, Y.; Li, J.; Li, W.; Jin, H.; Fang, F. Automatic Identification Technology of Lycium barbarum Flowering Period and Fruit Ripening Period Based on Faster R-CNN. *Chin. J. Agrometeorol.* **2020**, *41*, 668–677.
12. Ren, Y.; Du, Q. Fruit maturity recognition based on TensorFlow. *New Technol. New Prod. China* **2021**, *21*, 45–48.
13. Wang, T.; Zhao, Y.; Sun, Y.; Yang, R.; Han, Z.; Li, J. Recognition Approach Based on Data-balanced Faster R-CNN for Winter Jujube with Different Levels of Maturity. *Trans. Chin. Soc. Agric. Mach.* **2020**, *51*, 457–463, 492.
14. Liu, M.; Gao, T.; Ma, Z.; Song, Z.; Li, F.; Yan, Y. Target Detection Model of Corn Weeds in Field Environment Based on MSRCR Algorithm and YOLOv4-tiny. *Trans. Chin. Soc. Agric. Mach.* **2022**, *53*, 246–255, 335.
15. Xu, X. Research on APP for Rapid Detection of Rice Maturity Based on Computer Vision. Master's Theses, Jilin University, Changchun, China, 2021.
16. Bao, X.; Wang, S. Survey of object detection algorithm based on deep learning. *Transducer Microsyst. Technol.* **2022**, *41*, 5–9.
17. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
18. Zeng, G.; Yu, W.; Wang, R.; Lin, A. Research on Mosaic Image Data Enhancement and Detection Method for Overlapping Ship Targets. *Control. Theory Appl.* **2022**, 39, 1139–1148.
19. Yang, B.; Li, C.; Jiang, X.; Shi, H. A regularization loss function for improving the accuracy of deep learning classification models. *J. South-Cent. Univ. Natl.* **2020**, *39*, 74–78.
20. Wang, L.; Qin, M.; Lei, J.; Wang, X.; Tan, K. Blueberry maturity recognition method based on improved YOLOv4-Tiny. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 170–178.
21. Wan, L.; Ling, Y.; Zheng, X.; Li, X. Vehicle Type Recognition Based on MobileNet-YOLOv4. *Softw. Guide* **2021**, *20*, 173–178.
22. Zhang, F.; Chen, Z.; Bao, R.; Zhang, C.; Wang, Z. Recognition of dense cherry tomatoes based on improved YOLOv4-LITE lightweight neural network. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 270–278.
23. Wei, J.; Pan, S.; Tian, G.; Gao, W.; Sun, Y. Design and experiments of the binocular visual obstacle perception system for agricultural vehicles. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 55–63.
24. Gu, B.; Liu, Q.; Tian, G.; Wang, H.; Li, H.; Xie, S. Recognizing and locating the trunk of a fruit tree using improved YOLOv3. *Trans. Chin. Soc. Agric. Eng.* **2022**, *38*, 122–129.

25. Zhao, H.; Qiao, Y.; Wang, H.; Yue, Y. Apple fruit recognition in complex orchard environment based on improved YOLOv3. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 127–135.
26. Liu, Z. Visual Recognition and Maturity Detection Technology of Guava in Natural Environment. Master's Theses, South China Agricultural University, Guangzhou, China, 2019.
27. Xue, Y.; Huang, N.; Tu, S.; Mao, L.; Yang, A.; Zhu, X.; Yang, X.; Chen, P. Immature mango detection based on improved YOLOv2. *Trans. Chin. Soc. Agric. Eng.* **2018**, *34*, 173–179.
28. Long, J.; Zhao, C.; Lin, S.; Guo, W.; Wen, C.; Zhang, Y. Segmentation method of the tomato fruits with different maturities under greenhouse environment based on improved Mask R-CNN. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 100–108.
29. Zhang, Z.; Zhang, Z.; Li, J.; Wang, H.; Li, Y.; Li, D. Potato detection in complex environment based on improved YoloV4 model. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 170–178.
30. Chen, S.; Xiong, J.; Jiao, J.; Xie, Z.; Huo, Z.; Hu, W. Citrus fruits maturity detection in natural environments based on convolutional neural networks and visual saliency map. *Precis. Agric.* **2022**, *23*, 1515–1531. http://doi.org/10.1007/S11119-022-09895-2.
31. Subramanian, P.; Sankar, T.S. Detection of maturity stages of coconuts in complex background using Faster R-CNN model. *Biosyst. Eng.* **2021**, *202*, 119–132.
32. Luo, L.; Zou, X.; Ye, M.; Yang, Z.; Zhang, C.; Zhu, N.; Wang, C. Calculation and localization of bounding volume of grape for undamaged fruit picking based on binocular stereo vision. *Trans. Chin. Soc. Agric. Eng.* **2016**, *32*, 41–47.
33. Lei, W.; Lu, J. Visual positioning method for picking point of grape picking robot. *Jiangsu J. Agric. Sci.* **2020**, *36*, 1015–1021.
34. Luo, L.; Zou, X.; Xiong, J.; Zhang, Y.; Peng, H.; Lin, G. Picking Behavior of Grape Harvesting Robot Based on Visual Perception and Its Virtual Experiment. *Trans. Chin. Soc. Agric. Eng.* **2015**, *31*, 14–21.
35. Wang, J. Research on Vision Pre-positioning of Litchi Picking Robot under Large Field of View. Master's Theses, South China Agricultural University, Guangzhou, China, 2019.