agriculture

MDPI

*Article*

# Object Detection Algorithm for Lingwu Long Jujubes Based on the Improved SSD

**Yutan Wang *, Zhenwei Xing, Liefei Ma, Aili Qu and Junrui Xue**

School of Mechanical Engineering, Ningxia University, Yinchuan 750021, China
* Correspondence: wang_yt@nxu.edu.cn; Tel.: +86-139-9519-2065

**Abstract:** The detection of Lingwu long jujubes in a natural environment is of great significance for robotic picking. Therefore, a lightweight network of target detection based on the SSD (single shot multi-box detector) is presented to meet the requirements of a low computational complexity and enhanced precision. Traditional object detection methods need to load pre-trained weights, cannot change the network structure, and are limited by equipment resource conditions. This study proposes a lightweight SSD object detection method that can achieve a high detection accuracy without loading pre-trained weights and replace the Peleenet network with VGG16 as the trunk, which can acquire additional inputs from all of the previous layers and provide itself characteristic maps to all of the following layers. The coordinate attention module and global attention mechanism are added in the dense block, which boost models to more accurately locate and identify objects of interest. The Inceptionv2 module has been replaced in the first three additional layers of the SSD structure, so the multi-scale structure can enhance the capacity of the model to retrieve the characteristic messages. The output of each additional level is appended to the export of the sub-level through convolution and pooling operations in order to realize the integration of the image feature messages between the various levels. A dataset containing images of the Lingwu long jujubes was generated and augmented using pre-processing techniques such as noise reinforcement, light variation, and image spinning. To compare the performance of the modified SSD model to the original model, a number of experiments were conducted. The results indicate that the *mAP* (mean average precision) of the modified SSD algorithm for object inspection is 97.32%, the speed of detection is 41.15 fps, and the parameters are compressed to 30.37% of the original networks for the same Lingwu long jujubes datasets without loading pre-trained weights. The improved SSD target detection algorithm realizes a reduction in complexity, which is available for the lightweight adoption to a mobile platform and it provides references for the visual detection of robotic picking.

**Keywords:** long jujubes; target detection; SSD; convolutional neural network



**Citation:** Wang, Y.; Xing, Z.; Ma, L.; Qu, A.; Xue, J. Object Detection Algorithm for Lingwu Long Jujubes Based on the Improved SSD. *Agriculture* **2022**, *12*, 1456. https://doi.org/10.3390/agriculture12091456

Academic Editor: Koki Homma

Received: 29 July 2022
Accepted: 9 September 2022
Published: 13 September 2022



## 1. Introduction

Lingwu long jujubes are an economically important forest fruit in the Ningxia Hui Autonomous Region [1–3]. Lingwu long jujubes, also known as horse tooth jujubes, are long oval-shaped fruit, they are colorful and thick, and are rich in various vitamins and minerals. In particular, with the highest content of vitamin C, it is a unique variety of jujube in Ningxia [4]. There are some problems that limit the development of the Lingwu long jujubes industry, such as short harvesting cycles, high labor intensity, low picking efficiency during the harvesting period, time consuming and labor intensive, and pickers are prone to fatigue, which causes a lower harvesting productivity. However, robot picking can work in a continuous operation, which not only decrease the light workload and save manpower, but can also improve the harvesting productivity. Therefore, it is essential to study the detection of Lingwu long jujubes to boost the orientation accuracy of the picking robot.

The fast growth of computers and deep learning-based object inspection algorithms have been applied to many areas of life. In recent years, many results have been recorded

for object detection depth-based learning [5–13]. Kuznetsova et al. [14] proposed the YOLOv3-based DCNN (deep convolutional neural network) methods of calculation with pre-treatment and post-treatment techniques to achieve 90% precision and 19 ms detection speed for apple detection. Gao et al. [15] stated an apple detection method based on the Faster R-CNN (Region-Convolutional Neural Network), the *mAP* is 0.879, and the process of an image needs an average of 0.241 s. Lin et al. [16] adopted a Faster R-CNN based target detection framework for strawberry painting recognition, which was trained on 400 images by multi-layer processing images with an average precision of 86.1%. The method was designed for automatic and reliable strawberry flower inspection and delivers an essential tool to enable further study to be conducted to evaluate the strawberry yield in outdoor fields. Fu et al. [17] stated a kiwifruit field image detection technology based on the DCNN, using backwards propagation and SGD (stochastic gradient descent) techniques to coach the images, with an average accuracy of 89.3%, a time required to process each image of 0.274 s, and only 5 ms of the average time required to detect a fruit.

For the identification of the jujube, some scholars have also carried out some studies [18–23]. The above research studies are still not accurate enough for detecting and classifying jujubes, and the detection speed is slow. Qi et al. [24] established a machine vision system based on MIL9.0 and VB for Hami jujubes, which the identification rate can reached 83%. Ma et al. [25] made an equipment based on the machine vision for red jujube grading, for which the passing rate was superior to 80%. Li et al. [26] adopted an improved YOLOv5 detection algorithm based on the attention mechanism to optimize the loss function and post-treatment operation by introducing the attention mechanism of the Efficient Channel Attention and Coordinate Attention and the mean average precision reached 97.4%. Lu et al. [27] introduced a method based on the YOLOV3 model for the winter-jujube, the precision was 97.28% and the average time for each image was 1.39 s. These methods can obtain a good inspection accuracy, but the speed is slow.

In addition, all current target detection algorithms need to be loaded with pre-loaded training weights. This is because using pre-loaded training weights has advantages such as a better initialization performance and faster convergence after adding the pre-trained weights. Liang et al. [28] adopted the modified VGG16-based SSD network for mango detection, for which the precision was 0.920 with the detection speed was 35 fps. Xie et al. [29] made a pruning of the SSD model, which had the *mAP* of 76.31%. Zhao et al. [30] introduced a method for mutton based on the SSD model, which had the *mAP* of 93.07% and the inspection speed reached 10.52 fps. While these methods can fulfill the inspection task, the inspection performances are not excellent, they are time-consuming and the structure of the models are complex. However, there are some problems with loading pre-trained weights. For example, (1) the pre-trained model structure is large, the number of parameters is big, the network model structure has a poor flexibility, and it is difficult to change the network structure, which is intensive and complex, thereby limiting the application scenarios. (2) The loss functions and classification distributions of the detection tasks are different, and there are differences in the optimization space. (3) Although network fine-tuning can reduce the variability of the distribution of different target categories, the differences are too significant for the fine-tuning effect to be noticeable. The conventional SSD model requires a pre-loaded training weight to obtain a good accuracy performance, and the parameters are large, the calculation is complex, and the network architecture is tough to modify, which is easily impacted by the computational capacity of mobile devices, whereas the limited resources of the devices embedded in robot picking require a network with a few parameters and a lightweight structure. On the basis of the above existing problems, the study presents a lightweight SSD-based objection detection model that does not require pre-loaded training weights. The proposed model, a modified version of the original SSD, is simple and easy to modify, has fewer parameters, achieves a better detection accuracy, and has been used for the first time for the real-time robotic picking of Lingwu long jujubes.

## 2. Materials and Methods

### *2.1. Image Collection of Lingwu Long Jujubes*

The Lingwu long jujube images were collected between July and October 2021 in the Lingwu long jujubes base, in Lingwu City, Ningxia Hui Autonomous Region, China. The images were captured with a Samsung mobile. The dataset was made up of 4000 RGB (red, green, blue) images, the image quality resolution was 1706 × 1279 pixels, and the image format was jpg. The images were captured for image collection at the particular time period (7:00 until 11:00, 15:00 until 18:00), including the following conditions: different obstructions (leaves, branches), different light conditions (sunlight, sidelight), different backgrounds (soil, branch), shoot angle (below, lateral) and different weather (sunny, cloudy). Some Lingwu long jujubes sample images are seen in Figure 1.

(**01**) (**02**) (**03**) (**04**)

**Figure 1.** Sample images under different environments: (**01**) Occluded leaves; (**02**) Sunlight angle; (**03**) Soil background; (**04**) Leaves background.

### *2.2. Lingwu Long Jujubes Sample DataSets*

In the previous phase, only 4000 valid images of Lingwu long jujubes were collected, and a few training samples were not enough to train with good robustness. Therefore, the datasets were expanded by employing a label transformation with a strong generalization capability. If the training samples were too few, which would make the chosen samples inadequate to reflect the pre-classified requests and result in a poor generalization and over-fitting of the model. The data enlargement was utilized to randomly modify the training samples in order to obtain similar but not identical samples, and thereby extended the amount of the datasets, which could prevent a sample imbalance and improve the generalization capacity. Therefore, to avoid over-fitting the convolutional neural network to the training experimental data, different approaches of the data augmentation were introduced to enhance the model robustness. The images were subjected to noise reinforcement, illumination change, and image rotation. With the data enlargement, the number of image samples was broadened to 6000. Some Lingwu long jujube image samples after the data augmentation are seen in Figure 2.

(**01**) (**02**) (**03**) (**04**)

**Figure 2.** Images of Lingwu long jujubes after the data augmentation: (**01**) Add noise; (**02**) Boost brightness; (**03**) Decrease brightness; (**04**) Random rotation.

The manual annotation of image datasets is an essential part of the goal inspection algorithm. The PASCAL VOC2007 (Visual Object classes Challenge 2007) [31] dataset format is adopted in study, and the Labelimg 1.8.6 (Tzutalin, Vancouver, BS, Canada) tool is used to label the images of the Lingwu long jujubes, which can obtain the image in SSD format (x, y, w, h), that is, the center point, width, and height coordinates of the labeled image. The annotated Lingwu long jujube images are stored in xml format to obtain the Lingwu long jujube datasets with the equal images and the comment file names. The broadened datasets are randomly portioned into a training set and test set. The ratio of the two sample sets was approximately 8:2 [32], which not only guaranteed the random allocation of the datasets but also satisfied the validity of the assessment. By divided the dataset randomly, we can obtain the training set of 4800 and the test set of 1200.

### *2.3. Experimental Setup*

All tests are executed on the same PC and the computer configurations are as follows: Xeon-5118 CPU@ 2.3 GHz, 64 GB memory, NVIDIA TitanX GPU (Santa Clara, CA, USA), Ubuntu 16.04 system (Mark Shuttleworth, South Africa), Python version 3.6.13, PyTorch version 1.6.0, CUDA 10.0.130, and cuDNN 7.6.4 (Facebook AI Research, Menlo Park, CA, USA).

When training the SSD target detection model, the learning rate of the network training, the momentum factor, the decay rate of weight, and the batch size were set to 0.00025, 0.9, 0.0005, and 4, respectively. The networks were trained with a stochastic gradient descent (SGD).

### *2.4. Evaluation Indicators*

Four evaluation indicators are utilized to measure the efficiency of the claimed algorithm: average precision (*AP*), average recall (*AR*), number of network parameters, and detection speed [32]. The *AP* is the region composed of *P* (precision) and *R* (recall) as horizontal and vertical coordinates, which reflects the detection sensitivity of the model structure to the inspection object. *P* is utilized to evaluate the accuracy of the positive samples. *R* is the percentages of the accuracy of the overall test samples. The *AP* is an estimation exponent of the inspection performance for the overall test set. The *AR* is the IOU (intersection over union) from 0.5 to 0.95 and *R* is measured at an interval of 0.05 steps. Finally, the average is obtained. With only one target to detect, the *mAP* is the same as the *AP*. The expressions of the assessment indexes are as indicated:

$$mAP = AP = \int_0^1 P(R)dR \tag{1}$$

$$P = \frac{TP}{TP + FP} \times 100\% \tag{2}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{3}$$

where *TP* (True Positives) indicates that detected images was rightly identified, *FP* (False Positives) indicates that detected frames were misidentified Lingwu long jujubes, and *FN* (False Negatives) indicates that Lingwu long jujubes were misrecognized as with other matters.

### *2.5. Structure of the Peleenet Module*

The classical SSD target detection networks mainly use VGG16 as the backbone to obtain network feature maps and target classification tasks in order to achieve a more accurate and efficient target detection [33]. This paper aims to better and quickly extract the features of the Lingwu long jujube images and achieve a better detection accuracy even without loading the pre-trained models. On the basis of the traditional SSD object inspection network, the Peleenet network with an improved dense connectivity mechanism is used as the backbone network. Each layer of the neural network has only one direct connection to the next layer, and each layer obtains all of its preceding layers as additional inputs to enhance the reuse of the network for the image feature extraction. The feature

mapping of each layer is transmitted to all subsequent layers, and serves as the input to all subsequent layers, which strengthens the transmission of the network feature information, alleviates the gradient disappearance problem, and reduces the number of parameters to a great extent.

To make the backbone network have a good feature extraction effect and a simple structure network, the Peleenet network is improved in this paper [34]. The modified method structure is displayed in Figure 3, and the improvement methods are as follows:

- In contrast to the primal network construction, the modified model uses only the first two dense block modules [35]. The numbers of convolutional groups in the dense connection mechanism are 6 and 8, respectively, instead of 3 and 4 in the original network in order to deepen the network and improve the feature extraction capability of the Lingwu long jujubes images;
- The attention modules are added at the end of each convolutional group in the dense block module in order to suppress the unimportant network features while focusing the network more on the region of interest with the addition of an attention mechanism;
- The final pooling layer 2 × 2 convolution of this module is replaced to 3 × 3, its step size is altered to 1, and the obtained feature map is modified from 19 × 19 pixels to 38 × 38 pixels to satisfy the demand of the input feature map for the object detection network.

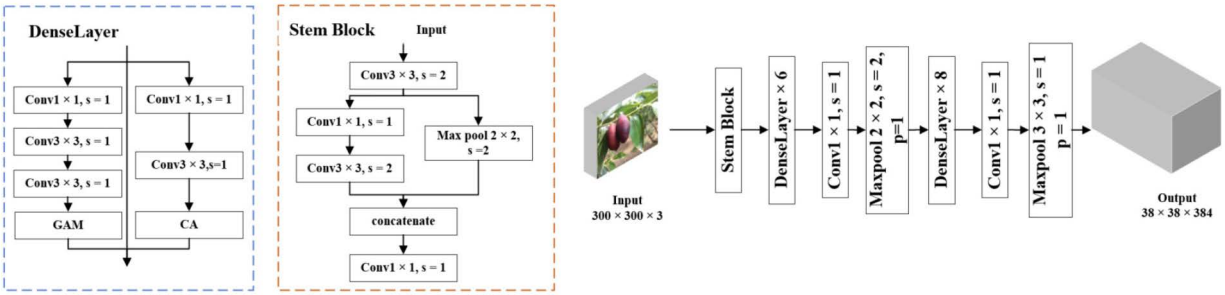


**Figure 3.** The structure of the improved Peleenet module. s indicates stride; p indicates padding.

### *2.6. CA Module and the GAM Module*

The CA (coordinate attention) module is a lightweight attention mechanism for network design [36]. The module structure is shown in Figure 4. The module first disintegrates the channel attention to make two 1D characteristics. The networks can capture long-distance attachments alongside one space dimension and retain the correct placement messages along the other orientation. The resulting characteristic maps are coded respectively to generate a couple of orientation-aware and position-sensitive characteristic maps. Finally, they are additionally employed to the export characteristic maps to boost the performance of the targets of choice.

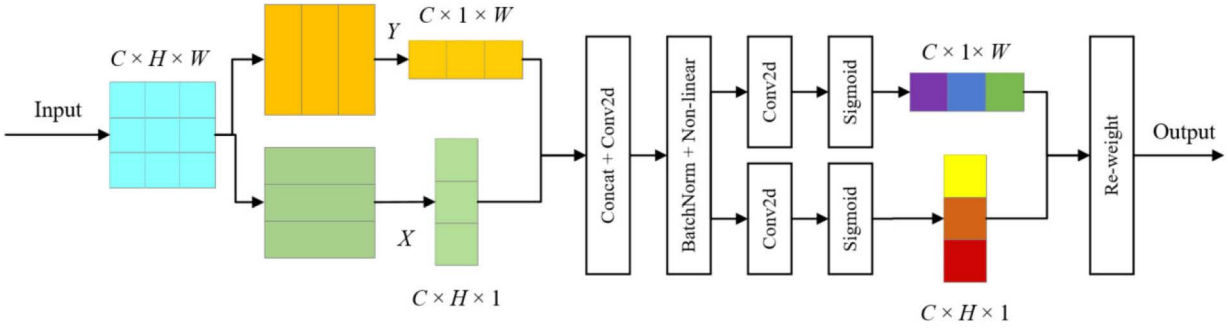


**Figure 4.** The construction of the CA module. *C* Indicates channel; *H* Indicates height; *W* Indicates width.

As Figure 4 shows, given an input $X$, the global averaging pooling operation is performed along the horizontal and vertical coordinates to encode each channel, and the channel attention is kept in the linear and upright orientations of the long-distance dependence. The $c$ export channel and $h$ can be written as shown in Equation (4):

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i \le W} x_c(h, i) \tag{4}$$

In the same way, the $c$ export channel and $w$ are shown in Equation (5):

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j \le H} x_c(j, w) \tag{5}$$

Then, the pooling operations in both directions are concatenated together for the convolution operation to interact with the information in both directions. Following the batch normalization and nonlinear activation function operations, this feature map is partitioned for the convolution operation separately. Finally, the sigmoid activation function is used.

The GAM (global attention mechanism) [37] module is an adaptation of the CBAM (convolutional block attention module) [38] and operates sequentially on the channel and space concentration. Yet the channel and space interactions are omitted, as well as the loss of the trans-dimension information. The GAM module can scale up the integral inter-dimensional characteristics with decreasing the spread of feature information. The module structure is shown in Figure 5.

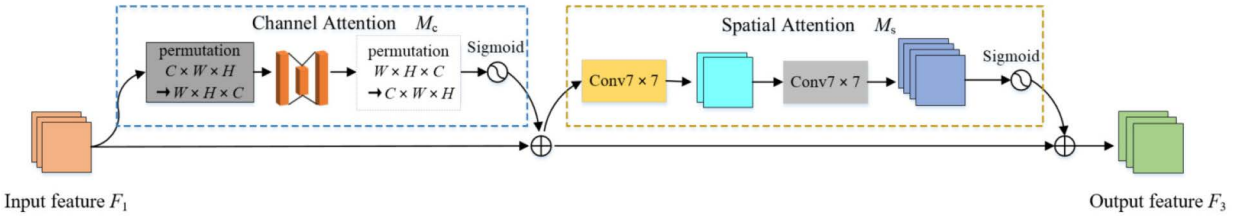


**Figure 5.** The construction of the GAM module.

Assuming that the entry of the characteristic maps $F_1 \in R^{C \times H \times W}$ and the output $F_3$ are shown as:

$$F_3 = M_s(M_c(F_1) \otimes F_1) \otimes (M_c(F_1) \otimes F_1) \tag{6}$$

where $M_c$ means the channel attention maps and $M_s$ means the space attention maps, $\otimes$ means multiply-by-element operation.

The channel module adopts a 3-dimensional arrangement to reserve the feature information in three dimensions, which employs a 2-layer MLP (multi-layer perception) to amplify the inter-dimensional channel-space reliance. The space attention adopts two convolutional layers and the scaling ratio $r$ for the space messages integration with the purpose of making the network more concerned with space information. To prevent decreasing the feature information to remaining characteristic maps without using the max-pooling operation.

### *2.7. Inceptionv2 Module*

The Inceptionv2 module is a multi-scale structured network module proposed by GoogLeNet [39], which is improved on the Inceptionv1 block. The Inceptionv2 block is featured in Figure 6. The adoption of two 3 × 3 convolutions instead of a 5 × 5 convolution in the original structure. Although a sizeable convolutional kernel can bring a larger perceptual field, it also brings more parameters to make the network computation more complex. However, the adoption of a small convolutional kernel instead of a large convolutional kernel maintains the network to obtain the same perceptual field while reducing the

number of network parameters and deepening the network. The BN (batch normalization) layer is introduced to normalize the export of each layer, thus, increasing the robustness of the network's model, making the network structure be trained at a more significant learning rate and the network to converge faster.

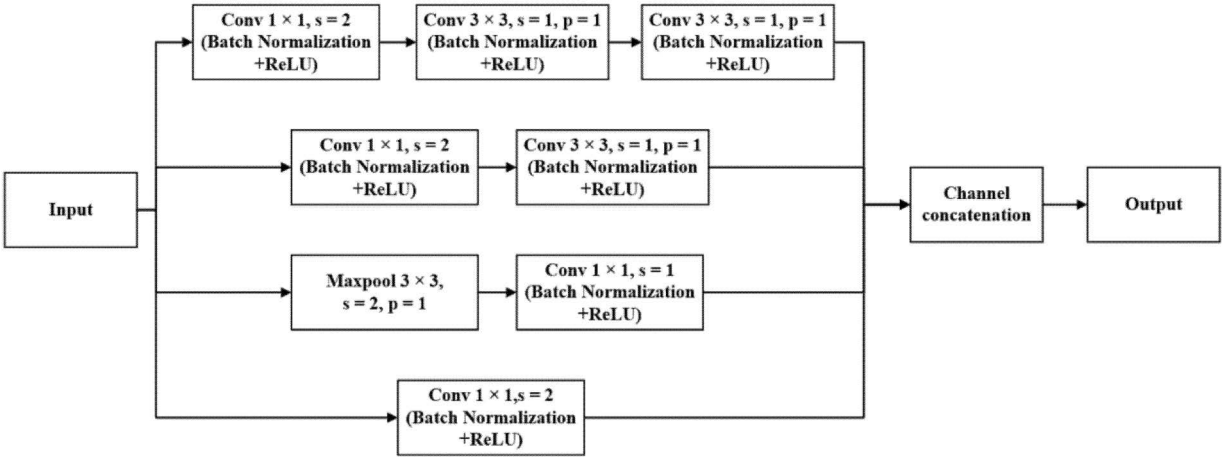


**Figure 6.** Inceptionv2 module.

### *2.8. Modified SSD Object Detection Structure*

In this paper, some improvements are made on the foundation of the traditional SSD target testing network to improve the network recognition of the detection targets. The modified network module architecture is described in Figure 7. Compared with the traditional SSD networks, the major improvements are as follows:

1. The VGG16 network in the original network is switched to the modified Peleenet module as the trunk network;
2. The convolution module in the first three additional layers of the original network is exchanged by the Inceptionv2 module. The multi-scale network structure of the module is utilized to enhance the network depth and further strengthen the capacity of the object detection network to retrieve the multi-scale messages from the jujubes;
3. The output of each additional level is appended to the export of the sub-level through the convolution and pooling operations to realize the integration of image feature messages between the various levels.

To meet the input image size of the SSD target detection network, all images of the Lingwu long jujubes are adjusted to 300 × 300 pixels as the export of the detection networks, and the characteristic map of 38 × 38 pixels × 384 channels is obtained after the image features are extracted through the Peleenet backbone network. The feature maps are acquired after the three Inceptionv2 feature extraction network modules are fused with the output of the features from the previous layers, and the feature maps of (19 × 19, 10 × 10, and 5 × 5) are obtained. Once the two additional layers of 1 × 1 convolutions and 3 × 3 convolutions are combined with the output of the previous network layer for the feature fusion, the feature maps of (3 × 3 and 1 × 1) are obtained separately. The six feature maps are subjected to the detection and classification operations. Then, overlapping and inaccurate detected bounding boxes are suppressed by the NMS (non-maximum suppression) algorithm in order to achieve the eventual target detection outcome.

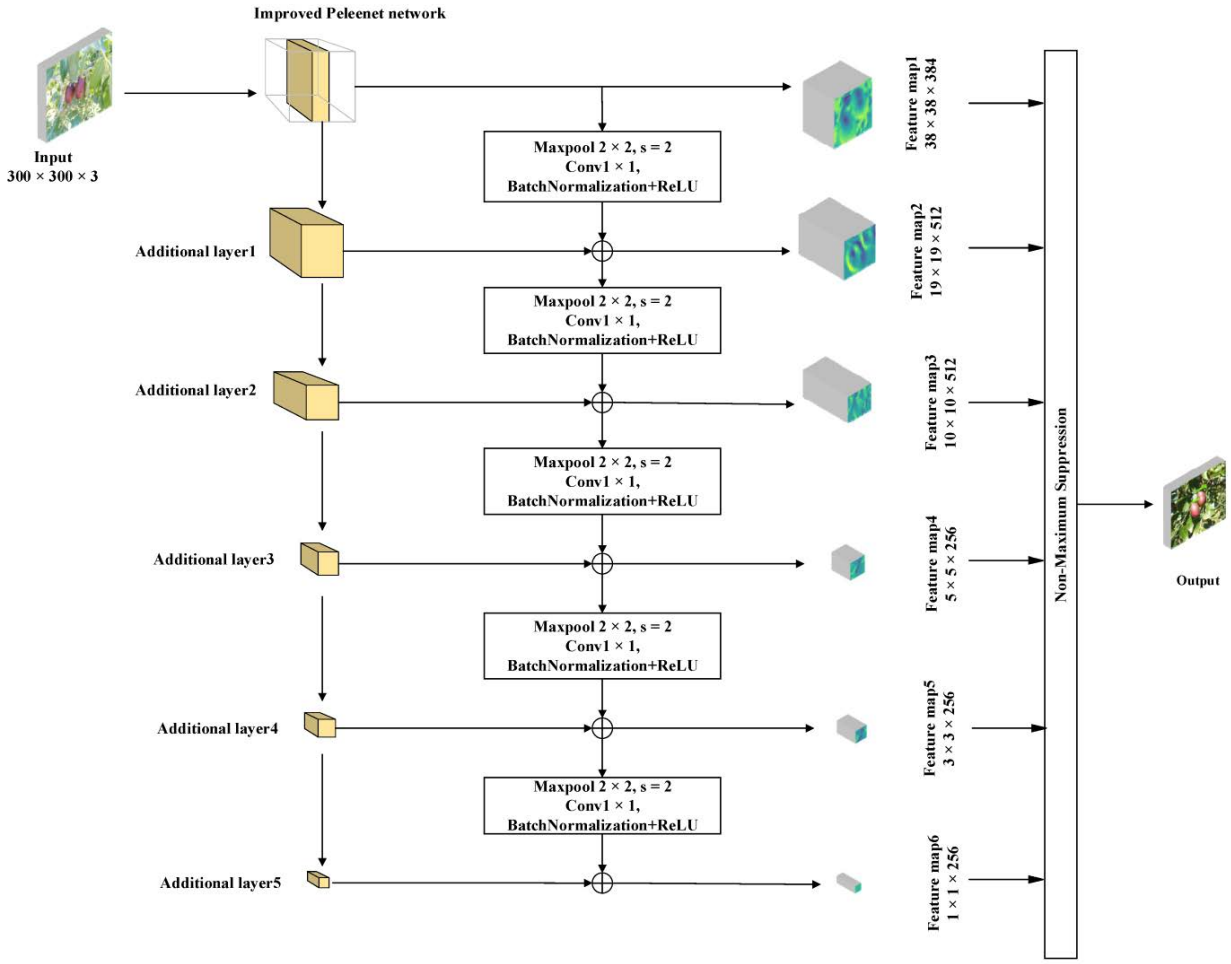


**Figure 7.** The improved SSD target detection model structure.

## 3. Results

To enhance the model detection performance, some modifications were performed on the conventional SSD model. The primary operations were not loading pre-trained training weights, substituting the trunk network, adding the attention module, exchanging the extra level and joining the multi-level integration. Four images of Lingwu long jujubes were randomly selected from the test sets, including an image with a small object, a regular image, a masked image and an image featuring jujubes of different sizes. The test images are shown in Figure 8.

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

**Figure 8.** Original image of the Lingwu long jujube test. (**a**) Image with a small object; (**b**) Regular image; (**c**) Masked image; (**d**) Image of jujubes of different sizes.

### *3.1. Comparison of the Improved Model and the Original Model*

To verify the validation of the model without pre-loaded training weights, the improved SSD model with the SSD model (using pre-loaded training weights) and the SSD model (without using pre-loaded training weights), the variation curves of the average precision and the loss value in the experimental comparison are displayed in Figure 9. The analysis outcomes of the Lingwu long jujube images are described in Figure 10. The evaluation indicator outcomes are given in Table 1.

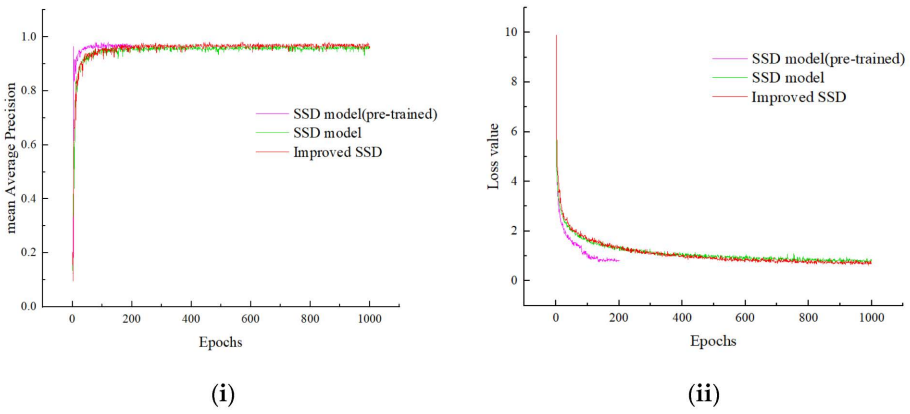


**Figure 9.** The change curve in the comparison test whether to load the pre-trained model. (**i**) Mean average precision. (**ii**) Loss value.

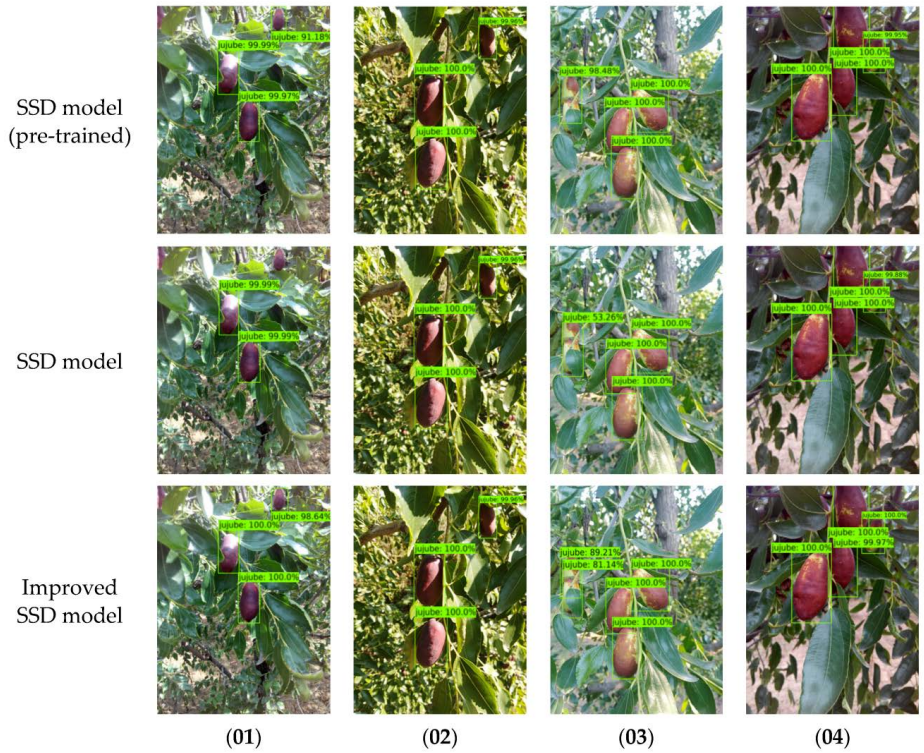


**Figure 10.** The comparison of the improved model and the original model. (**01–04**) The detection outcomes of the three models.

**Table 1.** Whether to load the pre-trained weights detection evaluation indicator results.

| Methods | Backbone | Pre-Trained Weights | CA + GAM | Inceptionv2 | Multilevel Fusion | mAP (%) | AR (%) | Speed (Fps) | Parameters/ ($\times 10^6$) |
|---|---|---|---|---|---|---|---|---|---|
| SSD model | VGG16 | √ | × | × | × | 97.19 | 76.81 | 45.39 | 11.92 |
| SSD model | VGG16 | × | × | × | × | 96.35 | 75.67 | 45.39 | 11.92 |
| Improved SSD model | Improved Peleenet | × | √ | √ | √ | 97.32 | 78.23 | 41.15 | 3.62 |

As shown in Figure 9, the improved SSD model and SSD model (without loading pre-trained weights) converge to stability after 800 epochs, while the SSD model (loaded with the pre-trained weights) converges quickly due to the loading pre-trained weights, and the network converges to stability after 200 epochs, but the mAP differs little from the improved SSD model. The mAP of the SSD model (without loading pre-trained weights) converges slightly more slowly than the improved SSD model due to not loading the pre-trained weights, and the mAP is 0.13% higher than that of the improved SSD model. When the network is trained enough, without using pre-trained weights, it can also achieve a good training accuracy.

The outcomes are shown in Figure 10, the SSD model (with the pre-loaded training weights) and the modified SSD model were good for detecting Lingwu long jujubes in the images. The SSD model (loaded with pre-trained weights) had a confidence level of 91.18% for the detection of the jujube number (a) in image (01). In image (03), two jujubes (b) and (c) were mistakenly detected as one, while the confidence level reached 98.48%. The SSD model (without pre-loaded training weights) was relatively poor in detection, which did not detect the jujube number (a) for image (01). In image (03), the Lingwu long jujubes numbers (b) and (c) were mistakenly detected as one with a confidence level of 53.26%.

As Table 1 shows, the improved SSD model with the Peleenet network compared with the VGG16 network is more complex and dense. Moreover, the extra layer introduction of the Inceptionv2 module is beneficial for the network to extract the multi-scale feature information, and the additional multi-level fusion is convenient for the network to integrate the multiple levels information before and after the network, which enriches the feature information of the network. Therefore, the inspection properties of the improved SSD module is outperforms the SSD module. The *mAP* of the modified SSD module is 0.97% superior to the SSD model. The *AR* of the modified SSD module is 2.56% over the SSD model. A slight difference in the *mAP* and the *AR* between the SSD model (loaded pre-trained weights) and the improved SSD model. The higher detection speed than the improved SSD model is 4.24 fps. However, the number of network parameters more than that of the improved SSD model is $8.3 \times 10^6$. Due to the limitation of equipment resource conditions, the network with the lightweight and higher detection accuracy is selected to fulfill the demands of on-site detection. Therefore, the modified SSD model is more amenable for adoption of the mobile machines than the SSD model.

### *3.2. Improved Peleenet Network Structures*

In this study, the CA module in the modified SSD structure is removed and marked as the SSD-A model, and the convolutional groups in the Peleenet network in the SSD model are altered as the SSD-B model with 3 and 4. The effectiveness of the improved Peleenet network is compared by using the same datasets with the experimental equipment. The variations in the average precision and loss value of the training results for each group are displayed in Figure 11.

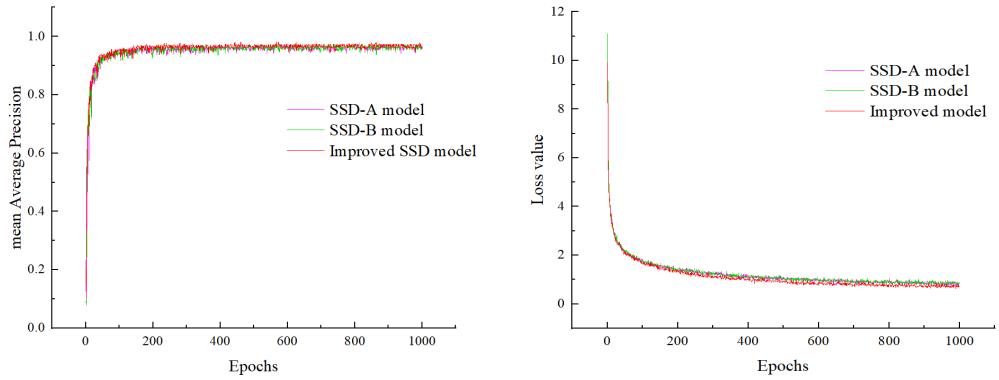


**Figure 11.** The change curve in the comparison experiment of the improved Peleenet network. (**i**) Mean average precision; (**ii**) Loss value.

The outcomes are shown in Figure 10, the *mAP* of the SSD-A model, SSD-B model, and improved SSD model all fluctuate in a small range, and the loss values fundamentally converge after 800 epochs, indicating that the network training has achieved a good training effect. By contrast with the improved SSD structure, the SSD-B backbone network structure adds more convolutional groups. During the initial training of the network, the convergence speed of the SSD-B model is excellent than the improved SSD architecture. By adding a number of iterations, the *mAP* of the modified SSD model is slightly superior to the SSD-A model, illustrating that the increasing network depth can make the network obtain a larger sensory field, capture similar features of more pixel points, and enhance the network's capacity to pick up target characteristics. The CA attention mechanism is introduced in the improved SSD model. When the network training is stable, the *mAP* of the improved SSD model outperforms the SSD-A model, and the curve volatility is smaller, which indicates the effectiveness of introducing the CA module in the improved SSD model.

By switching the attention mechanism and arranging the convolution group numbers, different modified models are acquired, respectively. The detection results of each improved target detection network for Lingwu long jujubes are exhibited in Figure 12. The network performance estimation comparison outcomes are given in Table 2.

**Table 2.** Improved Peleenet network comparison experiment evaluation indicators results.

| Methods | Backbone | Pre-Trained Weights | CA + GAM | Inceptionv2 | Multilevel Fusion | mAP (%) | AR(%) | Speed (Fps) | Parameters /($\times 10^6$) |
|---|---|---|---|---|---|---|---|---|---|
| SSD-A model | Improved Peleenet | × | × | √ | √ | 96.47 | 75.56 | 23.81 | 3.52 |
| SSD-B model | Improved Peleenet (3,4) | × | √ | √ | √ | 95.49 | 75.17 | 23.31 | 3.04 |
| Improved SSD model | Improved Peleenet | × | √ | √ | √ | 97.32 | 78.23 | 41.15 | 3.62 |

As shown in Figure 12, all of the improved models satisfactorily met the requirements of the detection task of the Lingwu long jujubes. Comparing the detection results, the SSD-B model is the least effective. For example, the jujube number (a) in image (01), the jujube numbers (b) and (c) were not detected. The SSD-A model had a confidence level of 70.87% for the jujube number (a) in image (01). The confidence level for the jujube (b) in image (03) was 53.27%, and jujube (c) was not detected, while the improved SSD model outperformed both the SSD-A and SSD-B models. The confidence level for the jujube number (a) in image (01) was 98.64%, and for the jujube numbers (b) and (c) in image (03), it was 89.21% and 81.14%, respectively. The validity of the improved Peleenet network is verified by the above experimental comparison.

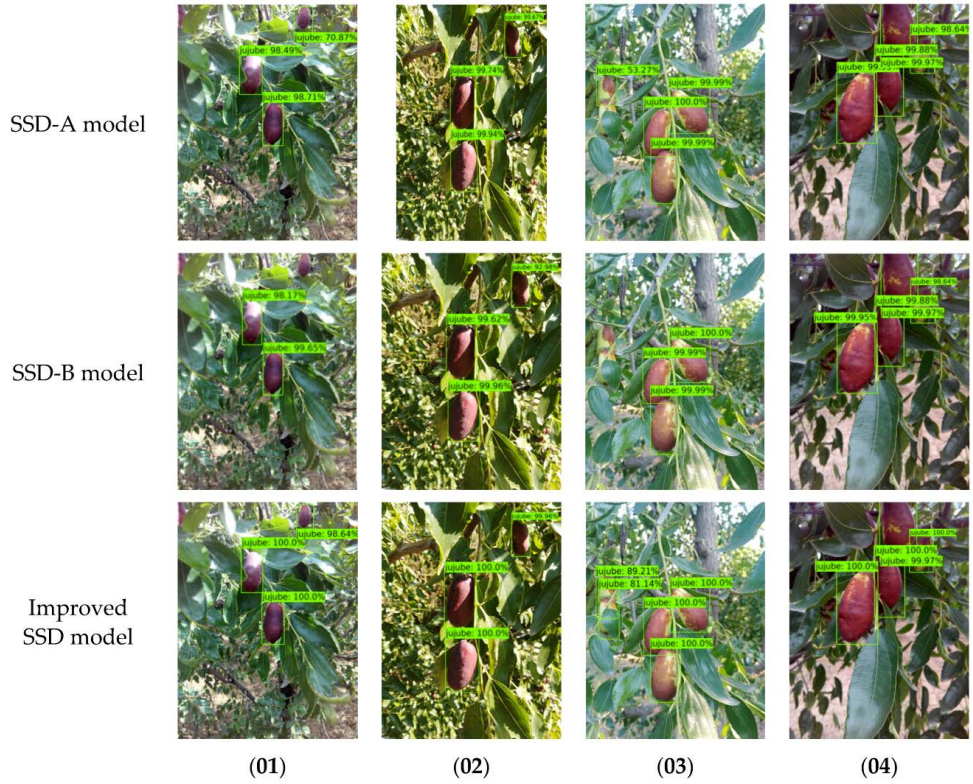


**Figure 12.** Improved Peleenet network contrast the experimental detection results. (**01–04**) The detection outcomes of the three models.

As Table 2 shows, the *mAP* of the modified SSD model is 0.85% superior to the SSD-A model and 1.83% greater than that of the SSD-B model. The *AR* of the improved SSD model is 2.67% and outperforms the SSD-A model and 3.06% superior to the SSD-B model. The reason for analyzing the high detection performance of the modified SSD structure is that the number of convolutional groups is increased in the dense block structure, which enhances the deepness of the network and improves the detection sensitivity. The CA module is introduced to reinforce the network characteristics extraction function. The efficiency of increasing the number of the convolutional groups and adding the CA attention mechanism is verified by comparison experiments.

### *3.3. Effectiveness of the Additional Layer Structure*

In this paper, the Inceptionv2 module structure in the improved SSD model is removed and marked as the SSD-C model, and the multi-level fusion structure is removed and marked as the SSD-D model. The variations in the average precision and loss values for each improved model structure are shown in Figure 13. The results of the target detection of the Lingwu long jujubes are shown in Figure 14, and the network performance assessment comparison outcomes are reported in Table 3.

As shown in Figure 13, the SSD-C model, SSD-D model, and the improved SSD model converge after reaching 800 epochs, the average precision tends to be stable, and the three networks achieve a good training effect. In contrast to the improved SSD structure, the *mAP* of both the SSD-C module and SSD-D model are inferior to the modified SSD model, and the modified SSD structure converges faster. The reason for the analysis is that the modified SSD model adds a multi-scale feature extraction module to avoid the gradient disappearance and speed up the convergence of the network and performs a larger nonlinear mapping to increase the expressiveness of the network. The introduction of a multi-level fusion module to facilitate the network fuses the multi-layer information and enriches the feature expression of the network. The comparison shows the effectiveness of adding the Inceptionv2 module, and multi-level fusion is verified.

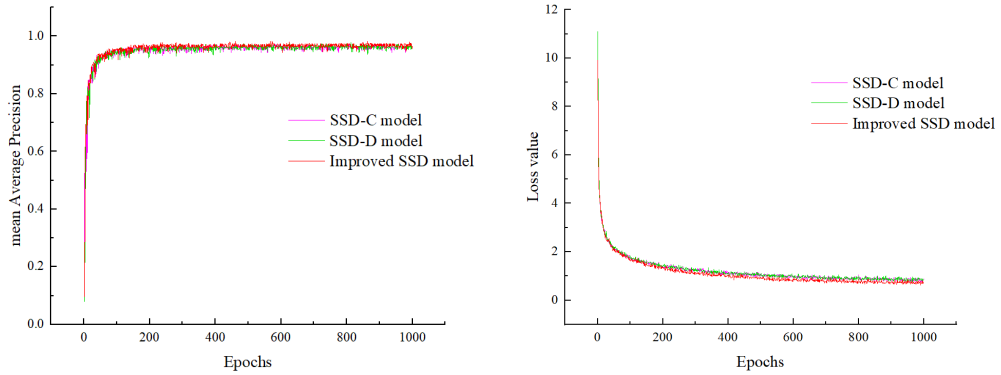


**Figure 13.** The average precision and loss value change curve in the comparison test of the improved SSD model structure. (**i**) mAP; (**ii**) Loss value.

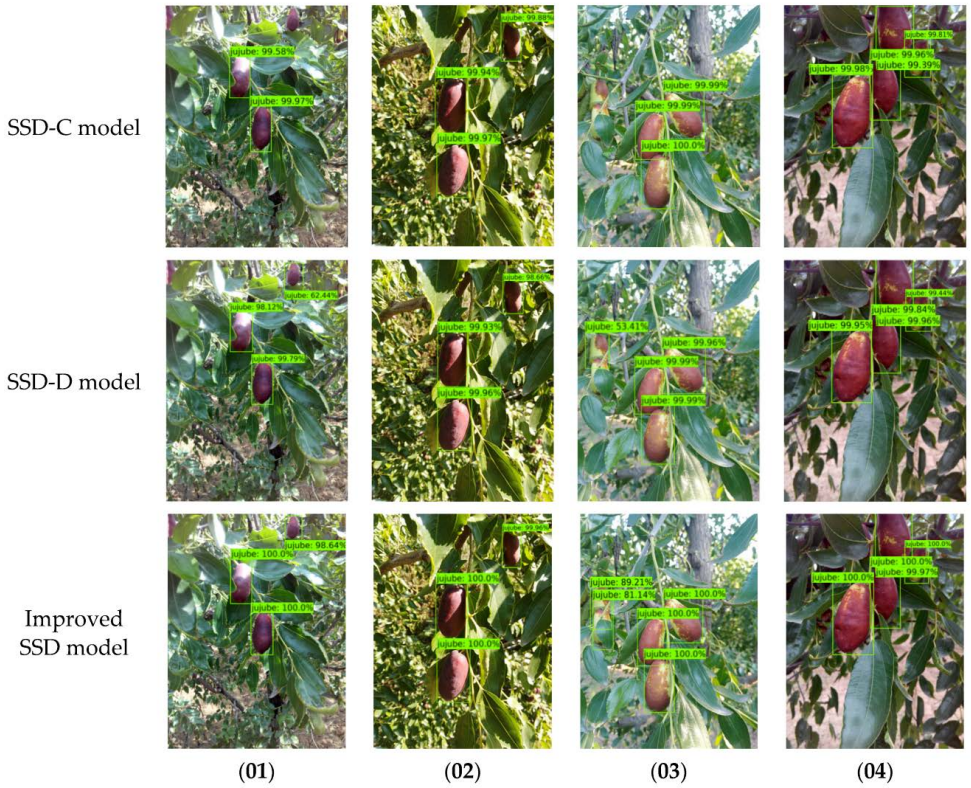


**Figure 14.** Improved SSD model structure comparison laboratory outcomes. (**01–04**) The detection outcomes of the three models.

**Table 3.** Results of the comparison experimental evaluation indicators of the improved SSD model structure.

| Methods | Backbone | Pre-Trained Weights | CA + GAM | Inceptionv2 | Multilevel Fusion | mAP(%) | AR(%) | Speed(Fps) | Parameters /($\times 10^6$) |
|---|---|---|---|---|---|---|---|---|---|
| SSD-C model | Improved Peleenet | × | √ | × | √ | 96.37 | 75.69 | 28.65 | 5.46 |
| SSD-D model | Improved Peleenet | × | √ | √ | × | 96.11 | 75.43 | 25.89 | 2.89 |
| Improved SSD model | Improved Peleenet | × | √ | √ | √ | 97.32 | 78.23 | 41.15 | 3.62 |

As seen in Figure 14, all of the improved models could achieve the detection of Lingwu long jujubes, but the SSD-C model had a less effective detection than the other models. For the jujube number (a) in image (01) and the jujube numbers (b) and (c) in image (03) were not detected. The SSD-D model had a confidence level of 62.44% for the jujube number (a) in image (03). The confidence level was 53.41% for the jujube number (b) in image (03), but the jujube number (c) was not detected. The improved SSD model outperformed the other two improved models in detecting each image, and it revealed that the improved SSD model had an excellent detection performance.

As displayed in Table 3, the *mAP* of the improved SSD model is 0.95% and 1.21% superior to the SSD-C module and SSD-D structure, respectively. The *AR* of the modified SSD model is 2.54% and 2.80% and they outperform the SSD-C structure and SSD-D module, respectively, which shows that the inspection property of the modified SSD model outperforms the SSD-C model and SSD-D model. Although the parameters of the modified SSD structure is $0.73 \times 10^6$ superior to the improved SSD-D structure, the modified SSD model reaches a detection speed of 41.15 fps. The above comparison outcomes explain the improved SSD model effectiveness.

## 4. Discussion

Lingwu long jujubes begin to be harvested in the middle of September every year. Harvest time is only 20 days and a large workforce is required. To develop the Lingwu long jujubes industry, it is crucial to achieve intelligent harvesting operations. Due to the small size of the jujubes, conventional machine vision methods are limited in order to achieve a good detection efficiency. Xia et al. [40] extracted a dry jujube image with machine vision and the precision was higher than 94.00%. Jiang et al. [41] introduced a method with machine vision for dried jujubes, and the accuracy reached 80–85%. Ma et al. [25] proposed a detecting system based on machine vision for Hami jujubes, and the accuracy was 91.43% and the average detecting time was 80 ms. The above models all adopted classical inspection methods, which can accomplish the detection missions, but the accuracy was poor and the testing speed was too slow, so it is very hard to apply the model to a robot.

Liu et al. presented a recognition method based on an improved YOLOv3 for detecting the winter jujube, the *mAP* was 82.01% and the testing time for each image was 0.0723 s. A modified Jujube testing model was proposed in this paper. The significant improvements include the backbone network replacement, the introduction of attention mechanisms and a multi-level convergence. Although the *mAP* of the modified model reached 97.32%, which has an excellent detection result for Lingwu long jujubes in a natural environment, the recognition accuracy of Lingwu long jujubes with a more serious obscuration is not high, and there is even the phenomenon of missing detection. Therefore, there are still some areas for improvement in this network.

Compared with other target detection algorithms, the proposed network structure does not need to load pre-trained weights. The network structure is relatively simple, the number of parameters is small, the model structure is flexible, and it is relatively easy to change the network. Subsequent studies should apply to other fruit detection methods and enhance the model generalization. To enhance the robustness of the network, it is necessary to acquire more images under different lighting and weather conditions to expand the datasets. To obtain high-resolution images, cameras with high pixel capabilities prevent image blurring and ghosting. Avoid affecting the network detection effect and reduce the omission ratio of these images. Subsequent studies should try to replace the network structure with a lightweight structure. Meanwhile, the accuracy of the network for detecting targets is still maintained.

## 5. Conclusions

On account of the problems of poor recognition effectiveness and the slow detection speed of Lingwu long jujubes in a natural environment, this paper proposes an improved SSD model that can achieve a high detection accuracy without loading pre-trained weights

for the target detection of jujubes. The color of mature jujubes is noticeably distinguished from the background. In the ideal cases, it is relatively easy to split mature Lingwu long jujubes. However, under actual natural conditions, segmenting color-based shape features in images is challenging due to uneven illumination and other factors. Zhang et al. [42] introduced a color image processing method for jujubes, for which the accuracy was 92.11%. Wang et al. [43] studied a method based on color space for Chinese dates and the successful rate was 87%. Xiao et al. [44] proposed a method based on the crackle color for jujubes, for which the precision rate was 89.7%. Zhao et al. [45] put forward one method based on SVM (support vector machine) and the HIS (hue, saturation, value) color for jujubes and correct rate was 96.2%. By contrast with traditional testing methods based on the characteristics of color and shape, the pattern not only reduces the effects of uneven light exposure but also automatically extracts image features. The model can meet the practical application of picking robots in the detection rate and scale of the model to enhance the harvest success rate.

By analyzing the experimental outcomes, the Peleenet network is more suitable for feature extraction of Lingwu long jujubes contrasted with the VGG16 structure. The *mAP* of the modified SSD reaches 97.32%, the *AR* reaches 78.23%, and the number of parameters is compressed to 30.37% of the original networks. Even without using pre-loaded training weights, the modified model can still gain a great detection reliability. The superiority of the modified structure, compared with the conventional SSD algorithm, is verified. In addition, the main reason for the incorrect recognition of jujubes is mainly the interference from the leaves. Although the improved SSD model has not had a high confidence in detecting the occluded Lingwu long jujubes, compared with the SSD model, it can reduce the omission ratio to some extent. The model can detect jujubes under natural circumstances and supply vital data for the picking operation of the robot.

However, the fact remains that there are some questions with the modified model. The accuracy rate needs to be improved for the seriously obscured Lingwu long jujubes. Subsequent work should attempt to improve the accuracy for the heavily obscured Lingwu long jujubes and include the practical application of a picking robot armed with the proposed detection model in a natural environment.

**Author Contributions:** Conceptualization and methodology, Y.W. and Z.X.; software, validation, formal and data analysis, Z.X.; writing-original draft preparation, L.M. and A.Q.; review and editing, supervision, Y.W. and J.X. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Science Foundation of Ningxia (Project No: 2020AAC03034).

**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. Song, L.H.; Cao, B. Effect of cover-cultivation on soil temperature and growth of Ziziphus jujuba Mill. 'Lingwu Changzao'. In Proceedings of the 29th International Horticultural Congress on Horticulture—Sustaining Lives, Livelihoods and Landscapes (IHC): 3rd International Jujube Symposium, Brisbane, Australia, 17–22 August 2014; pp. 89–92.
2. Yu, K.Q.; Zhao, Y.R.; Li, X.L.; Shao, Y.N.; Zhu, F.L.; He, Y. Identification of crack features in fresh jujube using Vis/NIR hyperspectral imaging combined with image processing. *Comput. Electron. Agric.* **2014**, *103*, 1–10. [CrossRef]
3. Chen, J.P.; Liu, X.Y.; Li, Z.G.; Qi, A.R.; Yao, P.; Zhou, Z.Y.; Dong, T.; Tsim, K.W.K. A Review of Dietary Ziziphus jujuba Fruit (Jujube): Developing Health Food Supplements for Brain Protection. *Evid. Based Complement. Altern. Med.* **2017**, *2017*, 3019568. [CrossRef] [PubMed]
4. Wu, L.G.; He, J.G.; Liu, G.S.; Wang, S.L.; He, X.G. Detection of common defects on jujube using Vis-NIR and NIR hyperspectral imaging. *Postharvest Biol. Technol.* **2016**, *112*, 134–142. [CrossRef]
5. Aquino, A.; Ponce, J.M.; Andujar, J.M. Identification of olive fruit, in intensive olive orchards, by means of its morphological structure using convolutional neural networks. *Comput. Electron. Agric.* **2020**, *176*, 105616. [CrossRef]

6. Wang, C.L.; Liu, S.C.; Wang, Y.W.; Xiong, J.T.; Zhang, Z.G.; Zhao, B.; Luo, L.F.; Lin, G.C.; He, P. Application of Convolutional Neural Network-Based Detection Methods in Fresh Fruit Production: A Comprehensive Review. *Front. Plant Sci.* **2022**, *13*, 868745. [CrossRef] [PubMed]
7. Gene-Mola, J.; Sanz-Cortiella, R.; Rosell-Polo, J.R.; Morros, J.R.; Ruiz-Hidalgo, J.; Vilaplana, V.; Gregorio, E. Fruit detection and 3D location using instance segmentation neural networks and structure-from-motion photogrammetry. *Comput. Electron. Agric.* **2020**, *169*, 105165. [CrossRef]
8. Mai, X.C.; Zhang, H.; Jia, X.; Meng, M.Q.H. Faster R-CNN With Classifier Fusion for Automatic Detection of Small Fruits. *IEEE Trans. Autom. Sci. Eng.* **2020**, *17*, 1555–1569. [CrossRef]
9. Paturkar, A.; Sen Gupta, G.; Bailey, D. Apple Detection for Harvesting Robot Using Computer Vision. *Helix* **2018**, *8*, 4370–4374. [CrossRef]
10. Silwal, A.; Karkee, M.; Zhang, Q. A hierarchical approach to apple identification for robotic harvesting. *Trans. ASABE* **2016**, *59*, 1079–1086. [CrossRef]
11. Zhang, Y.C.; Yu, J.Y.; Chen, Y.; Yang, W.; Zhang, W.B.; He, Y. Real-time strawberry detection using deep neural networks on embedded system (rtsd-net): An edge AI application. *Comput. Electron. Agric.* **2022**, *192*, 106586. [CrossRef]
12. Yu, Y.; Zhang, K.L.; Liu, H.; Yang, L.; Zhang, D.X. Real-Time Visual Localization of the Picking Points for a Ridge-Planting Strawberry Harvesting Robot. *IEEE Access* **2020**, *8*, 116556–116568. [CrossRef]
13. Rodriguez, J.P.; Corrales, D.C.; Aubertot, J.N.; Corrales, J.C. A computer vision system for automatic cherry beans detection on coffee trees. *Pattern Recognit. Lett.* **2020**, *136*, 142–153. [CrossRef]
14. Kuznetsova, A.; Maleva, T.; Soloviev, V. Using YOLOv3 Algorithm with Pre- and Post-Processing for Apple Detection in Fruit-Harvesting Robot. *Agronomy* **2020**, *10*, 1016. [CrossRef]
15. Gao, F.F.; Fu, L.S.; Zhang, X.; Majeed, Y.; Li, R.; Karkee, M.; Zhang, Q. Multi-class fruit-on-plant detection for apple in SNAP system using Faster R-CNN. *Comput. Electron. Agric.* **2020**, *176*, 105634. [CrossRef]
16. Lin, P.; Lee, W.S.; Chen, Y.M.; Peres, N.; Fraisse, C. A deep-level region-based visual representation architecture for detecting strawberry flowers in an outdoor field. *Precis. Agric.* **2020**, *21*, 387–402. [CrossRef]
17. Fu, L.S.; Feng, Y.L.; Majeed, Y.; Zhang, X.; Zhang, J.; Karkee, M.; Zhang, Q. Kiwifruit detection in field images using Faster R-CNN with ZFNet. In Proceedings of the 6th International-Federation-of-Automatic-Control (IFAC) Conference on Bio-Robotics (BIOROBOTICS), Beijing, China, 13–15 July 2018; pp. 45–50.
18. Wang, Y.T.; Dai, Y.P.; Xue, J.R.; Liu, B.H.; Ma, C.H.; Gao, Y.Y. Research of segmentation method on color image of Lingwu long jujubes based on the maximum entropy. *EURASIP J. Image Video Process.* **2017**, *2017*, 34. [CrossRef]
19. Yuan, R.R.; Liu, G.S.; He, J.G.; Wan, G.L.; Fan, N.Y.; Li, Y.; Sun, Y.R. Classification of Lingwu long jujube internal bruise over time based on visible near-infrared hyperspectral imaging combined with partial least squares-discriminant analysis. *Comput. Electron. Agric.* **2021**, *182*, 106043. [CrossRef]
20. Geng, L.; Xu, W.L.; Zhang, F.; Xiao, Z.T.; Liu, Y.B. Dried Jujube Classification Based on a Double Branch Deep Fusion Convolution Neural Network. *Food Sci. Technol. Res.* **2018**, *24*, 1007–1015. [CrossRef]
21. Al-Saif, A.M.; Abdel-Sattar, M.; Aboukarima, A.M.; Eshra, D.H. Identification of Indian jujube varieties cultivated in Saudi Arabia using an artificial neural network. *Saudi J. Biol. Sci.* **2021**, *28*, 5765–5772. [CrossRef]
22. Feng, L.; Zhu, S.S.; Zhou, L.; Zhao, Y.Y.; Bao, Y.D.; Zhang, C.; He, Y. Detection of Subtle Bruises on Winter Jujube Using Hyperspectral Imaging With Pixel-Wise Deep Learning Method. *IEEE Access* **2019**, *7*, 64494–64505. [CrossRef]
23. Luo, X.Z.; Ma, B.X.; Wang, W.X.; Lei, S.Y.; Hu, Y.Y.; Yu, G.W.; Li, X.Z. Evaluation of surface texture of dried Hami Jujube using optimized support vector machine based on visual features fusion. *Food Sci. Biotechnol.* **2020**, *29*, 493–502. [CrossRef]
24. Qi, X.X.; Ma, B.X.; Xiao, W.D. On-Line Detection of Hami Big Jujubes' Size and Shape Based on Machine Vision. In Proceedings of the 2011 International Conference on Computer Distributed Control and Intelligent Environmental Monitoring, Changsha, China, 19–20 February 2011.
25. Ma, B.; Qi, X.; Wang, L.; Zhu, R.; Chen, Q.; Li, F.; Wang, W. Size and defect detection of Hami Big Jujubes based on computer vision. *Adv. Mate. Res.* **2012**, *562–564*, 750–754. [CrossRef]
26. Li, S.L.; Zhang, S.J.; Xue, J.X.; Sun, H.X.; Ren, R. A Fast Neural Network Based on Attention Mechanisms for Detecting Field Flat Jujube. *Agriculture* **2022**, *12*, 717. [CrossRef]
27. Lu, Z.; Zhao, M.; Luo, J.; Wang, G.; Wang, D. Design of a winter-jujube grading robot based on machine vision. *Comput. Electron. Agric.* **2021**, *186*, 106170. [CrossRef]
28. Liang, Q.; Zhu, W.; Long, J.; Wang, Y.; Sun, W.; Wu, W. A real-time detection framework for on-tree mango based on SSD network. In Proceedings of the 11th International Conference on Intelligent Robotics and Applications, Newcastle, NSW, Australia, 9–11 August 2018; pp. 423–436.
29. Xie, X.; Han, X.; Liao, Q.; Shi, G. Visualization and pruning of SSD with the base network VGG16. In Proceedings of the 2017 International Conference on Deep Learning Technologies, Chengdu, China, 2–4 June 2017; pp. 90–94.
30. Zhao, S.D.; Hao, G.Z.; Zhang, Y.C.; Wang, S.C. A real-time classification and detection method for mutton parts based on single shot multi-box detector. *J. Food Process Eng.* **2021**, *44*, e13749. [CrossRef]
31. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]

32. Yuan, T.; Lv, L.; Zhang, F.; Fu, J.; Gao, J.; Zhang, J.X.; Li, W.; Zhang, C.L.; Zhang, W.Q. Robust Cherry Tomatoes Detection Algorithm in Greenhouse Scene Based on SSD. *Agriculture* **2020**, *10*, 160. [CrossRef]
33. Sunil, G.; Zhang, Y.; Koparan, C.; Ahmed, M.R.; Howatt, K.; Sun, X. Weed and crop species classification using computer vision and deep learning technologies in greenhouse conditions. *J. Agric. Food Res.* **2022**, *9*, 100325. [CrossRef]
34. Wang, R.J.; Li, X.; Ling, C.X. Pelee: A Real-Time Object Detection System on Mobile Devices. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 2–8 December 2018.
35. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
36. Hou, Q.B.; Zhou, D.Q.; Feng, J.S. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13708–13717.
37. Liu, Y.; Shao, Z.; Hoffmann, N.J.A. Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions. *arXiv* **2021**, arXiv:2112.05561.
38. Woo, S.H.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
39. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
40. Xia, Y.; Chen, B.; Li, Y.H.; Chen, M. Application research based on region of the image threshold segmentation algorithm of RGB Jujube. *Mod. Instrum.* **2022**, *222*, 156–176.
41. Jiang, J.X.; Zhou, J.H. Dried Jujubes Online Detection Based on Machine Vision. *Adv. Mater. Res.* **2013**, *655–657*, 673–678. [CrossRef]
42. Zhang, J.X.; Ma, Q.Q.; Li, W.; Xiao, T.T. Feature extraction of jujube fruit wrinkle based on the watershed segmentation. *Int. J. Agric. Biol. Eng.* **2017**, *10*, 165–172. [CrossRef]
43. Wang, F.J.; Dong, Y.Q. Surface Defect Detection of Chinese Dates Based on Machine Vision. *Adv. Mater. Res.* **2011**, *403–408*, 1356–1359. [CrossRef]
44. Xiao, A.L.; Huang, X.C.; Zhang, J.X.; Li, W. The research of detecting method on crackled Chinese date based on chrominance components. *Biotechnol. Indian J.* **2014**, *10*, 4945–4954.
45. Zhao, J.; Liu, S.; Zou, X.; Shi, J.; Yin, X. Recognition of defect Chinese dates by machine vision and support vector machine. *Nongye Jixie Xuebao Trans. Chin. Soc. Agric. Mach.* **2008**, *39*, 113–116.