

Article

Tomato Young Fruits Detection Method under Near Color Background Based on Improved Faster R-CNN with Attention Mechanism

Peng Wang^{1,2,3}, Tong Niu^{1,2,3} and Dongjian He^{1,2,3,*} 

¹ College of Mechanical and Electronic Engineering, Northwest A&F University, Xianyang 712100, China; wpeng@nwsuaf.edu.cn (P.W.); niutong@nwafu.edu.cn (T.N.)

² Key Laboratory of Agricultural Internet of Things, Ministry of Agriculture and Rural Affairs, Xianyang 712100, China

³ Shaanxi Key Laboratory of Agricultural Information Perception and Intelligent Services, Xianyang 712100, China

* Correspondence: hdj168@nwsuaf.edu.cn; Tel.: +86-029-87092391

Abstract: The information of tomato young fruits acquisition has an important impact on monitoring fruit growth, early control of pests and diseases and yield estimation. It is of great significance for timely removing young fruits with abnormal growth status, improving the fruits quality, and maintaining high and stable yields. Tomato young fruits are similar in color to the stems and leaves, and there are interference factors, such as fruits overlap, stems and leaves occlusion, and light influence. In order to improve the detection accuracy and efficiency of tomato young fruits, this paper proposes a method for detecting tomato young fruits with near color background based on improved Faster R-CNN with an attention mechanism. First, ResNet50 is used as the feature extraction backbone, and the feature map extracted is optimized through Convolutional Block Attention Module (CBAM). Then, Feature Pyramid Network (FPN) is used to integrate high-level semantic features into low-level detailed features to enhance the model sensitivity of scale. Finally, Soft Non-Maximum Suppression (Soft-NMS) is used to reduce the missed detection rate of overlapping fruits. The results show that the mean Average Precision (mAP) of the proposed method reaches 98.46%, and the average detection time per image is only 0.084 s, which can achieve the real-time and accurate detection of tomato young fruits. The research shows that the method in this paper can efficiently identify tomato young fruits, and provides a better solution for the detection of fruits with near color background.

Keywords: convolutional neural network; feature pyramid network; near color background; tomato young fruits; fruit detection



Citation: Wang, P.; Niu, T.; He, D. Tomato Young Fruits Detection Method under Near Color Background Based on Improved Faster R-CNN with Attention Mechanism. *Agriculture* **2021**, *11*, 1059. <https://doi.org/10.3390/agriculture11111059>

Academic Editor: Maciej Zaborowicz

Received: 14 September 2021

Accepted: 22 October 2021

Published: 28 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The real-time information acquisition and monitoring growth status of tomato young fruits can grasp the early quality information of young fruits, and it is of great significance to timely remove abnormal fruits having deformities, diseases, insects, etc., to ensure the normal growth of healthy fruits and to improve fruit quality and yield [1]. The automatic monitoring of the growth status of tomato young fruits is an important part of the tomato production process. Therefore, it is necessary to develop agricultural robots and other technical means to complete this work. At present, most robots which used to complete advanced agricultural tasks, such as automatic flower thinning, fruit thinning and picking, are using object detection techniques to achieve fruits positioning, and the image-based object detection algorithm is the key factor affecting the recognition performance of robots [2–4]. Therefore, high-performance object detection algorithms play a fundamental role in improving the performance of robots, and they can provide theoretical guidance for the object recognition of fruit-thinning robots.

The traditional fruits detection method mainly extracts the color, shape, texture and other shallow feature information of the image and scholars have conducted extensive research on it [5–7]. Zhao et al. extracted the Haar-like features of tomato grayscale images and used the Ada Boost classifier to identify the fruits, and eliminated false positives in the classification results through the color analysis method based on Average Pixel Value (APV) [8]. This method used the combination of Ada Boost classification and color analysis to correctly detect 96% of mature tomatoes, but the false negative is approximately 10%, and 3.5% of tomatoes are not detected, which is seriously affected by the background. In order to reduce the interference caused by a complex background and lighting, Zhao et al. extracted two new feature images from $L^*a^*b^*$ color space and luminance, in-phase, quadrature-phase (YIQ) color space, respectively, a^* -component image and I-component image, used wavelet transform to fuse two feature images at the pixel level and then segment them to obtain tomato fruit recognition results [9]. In order to simplify the calculation and improve recognition efficiency, Wu et al. first extracted different texture features and color component information in each image block, using the Iterative RELIEF (I-RELIEF) algorithm to analyze related features and their weights, and then they used the weighted Relevance Vector Machine (RVM) classifier of selected related features to divide the image blocks into different categories, and finally obtained the fruits recognition results [10]. The above studies have improved the recognition accuracy of mature fruits. However, due to the small size of tomato young fruits, their color is similar to the stems and leaves, and there are effects such as stems and leaves occlusion and changes in ambient light. The method mentioned above is not sensitive to small-sized, near color background tomato young fruits, therefore it is difficult to achieve a stable recognition effect. Yamamoto et al. used a regression tree classifier to build a decision tree, extracted tomato fruit pixels through pixel segmentation, and detected single mature and immature tomato fruits in the image with an accuracy of 88% [11]. This method has a certain improvement in the ability of green tomato fruits recognition, but it has higher requirements for image quality and poor adaptability to noise. In summary, traditional fruits detection methods are mainly based on shallow features such as color and shape outline, and the adaptability to the changeable fruit morphology and tomato stems and leaves occlusion is not strong [12]. Therefore, it is difficult to achieve high-precision and real-time requirements for the recognition of tomato young fruits with near color background using traditional methods.

In recent years, with the rapid development of deep learning techniques and continuous improvement of computer computing power, Convolutional Neural Network (CNN) has shown great advantages in the field of object detection [13,14]. In the field of agriculture, compared with traditional fruit detection methods, CNN has better performance in tasks, such as image classification [15,16], object detection [17–20] and object segmentation [21,22]. Chen et al. proposed a dual-path feature extraction network to extract the semantic feature information of small tomato objects using the K-means++ clustering method to calculate the scale of the bounding box, and the test accuracy was up to 94.29% [23]. The studies mentioned above have made targeted improvements to the model in terms of different scales, environmental interference, and background removal and achieved good results. Wang et al. proposed a Region-Based Fully Convolutional Network (R-FCN) apple young fruits object detection method, which can effectively identify occluded, blurred and shadowed young fruits objects [24]. However, the problem of fruits overlap caused by the dense distribution of clustered fruits will lead to a large false detection and missed detection, which makes the generalization ability of the model insufficient and the recognition accuracy is relatively low [25,26]. At the same time, due to the presence of complex backgrounds in the orchard and the irregular growth status and growth position of tomato young fruits, it is more severely affected by the occlusion of tomato stems and leaves during the detection process.

In response to the above problems, this paper proposes a method for detecting tomato young fruits in a near color background based on improved Faster R-CNN with an attention mechanism. First, the pre-trained weights of the feature extraction network ResNet50 is

used and fine-tuned. In order to solve the problem that the feature is difficult to extract due to the occlusion of stems and leaves, the Convolutional Block Attention Module (CBAM) attention mechanism [27] is used to process the feature map to strengthen the regional characteristics of the fruits and increase the richness of the feature map. In addition, in order to enhance the model's adaptability to tomato young fruits of different scales, Feature Pyramid Networks (FPN) [28], which has a lower computational cost, is used to fuse high-level semantic features with low-level detailed features. Then, according to the growth characteristics of the young fruit clusters, the Soft Non-Maximum Suppression (Soft-NMS) method is used to reduce the missed detection rate of overlapping fruits. Finally, the Region of Interest Align (RoI Align) region feature mapping method is used to optimize the positioning of the bounding box, and the detection model of tomato young fruits is constructed.

The rest of the work is arranged as follows: In Materials and Methods, the source and structure of the data set used in this research are introduced, the improvement of related structure and algorithm of Faster R-CNN, as well as model training and testing, are discussed in detail. Results and Discussion presents the test to evaluate the performance of the model and analyzes the test results. The Conclusion summarizes the work of this paper.

2. Materials and Methods

2.1. Data Sources

RGB images contain only three channel information of red, green and blue, data redundancy is small and the cost of image acquisition is low, so this work uses RGB images of tomato young fruits. From April 2021 to May 2021, a total of 2235 images of tomato young fruits were collected in the agricultural digital greenhouse of Northwest A&F University to construct the data set. In the image acquisition process, taking into account the difference in imaging results caused by different weather and acquisition times, in order to ensure the diversity and effectiveness of the data set, images were collected on tomatoes transplanted for 30 days and with a fruit size of approximately 28 mm–55 mm in different weather conditions (sunny, cloudy) and different time periods (morning, noon, evening). The digital image acquisition device is an MI 9 smartphone, and the image size is 3000×3000 pixels. Because the clusters of densely distributed tomato young fruits have different growth status and positions, and there are stems and leaves occluded, etc., the fruits are photographed from different angles and directions to increase the diversity and complexity of the samples. Table 1 shows sample images taken under different conditions.

Table 1. Example images acquired under different conditions.

	Morning	Noon	Evening
Sunny			
Cloudy			

Image labeling is an important part of the object detection process. LabImage labeling software is used to annotate the position information of the tomato in the image. Annotate according to the format of PASCAL VOC 2007 data set (a data set containing multiple types of target labeled images and annotated files), and automatically generate

the corresponding XML file. Take any image in the data set as an example, Figure 1a,b are the position of the manual annotation box and the corresponding XML description file, respectively.

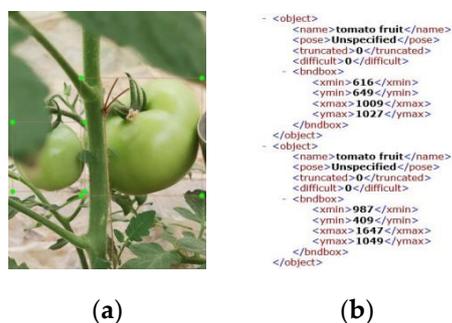


Figure 1. Annotation of tomato young fruits image. (a) annotated image. (b) XML document.

In order to train and test the object detection model, the annotated data set is randomly divided into 3 independent subsets: training set, validation set and test set. Divide 80% of the images in the data set into the training set, 10% into the validation set, and the remaining 10% into the test set to ensure that each subset contains different forms of fruit images. The training set is used to train the network, and each structure of the network is automatically learned by adjusting the weights and bias. The validation set is used to make a preliminary evaluation of the model and visually demonstrate the effect of model training through the recognition accuracy. The test set is used to evaluate the generalization ability of the model.

2.2. Feature Extraction Network

The PASCAL VOC 2007 data set of the original Faster R-CNN used for training and testing contains 21 different types of objects, and the feature differences among different types of objects are obvious. The VGG16 network used in the model can achieve good results in the feature extraction process. However, there are still some problems in detecting images of tomato young fruits collected in a field environment. First, the color of young tomato fruits is green, which is similar to the color of the tomato stems and leaves. At the same time, the occlusion of fruits by the stems and leaves in the images taken at different angles are also different. Second, the images acquired in the field environment all contain a large number of irrelevant and complex backgrounds, and there are environmental interference factors such as lighting. In addition, young fruits in the field are mostly clustered and densely distributed, with different growth status and uneven distribution. The traditional feature extraction network VGG16 has insufficient feature richness for the extraction of young tomato fruits with the above characteristics, and it is difficult to achieve satisfactory results.

In order to solve the problems mentioned above in the task of young tomato fruits detecting and improve the richness of feature extraction of young tomato fruits by the network, this paper adopts ResNet50 with residual structure as the feature extraction network. Figure 2 shows 2 different basic structures of ResNet50. The first layer structure of each residual body in the ResNet50 network structure is shown in Figure 2a. The input feature is extracted through the main branch and the depth of the input feature matrix is expanded to twice the input. The Shortcut branch uses 1×1 convolution to increase the dimension and adds it to the output of the main branch. The subsequent layer structure of each residual body is shown in Figure 2b. The main branch performs feature extraction, while Shortcut does not do any processing and directly adds to the output of the main branch, and the network learns the difference between the 2 branches. Due to the small sample size of acquired images, it is difficult to retrain the model to achieve good results. In order to avoid over-fitting problems, the transfer learning method is used to fine-tune the network according to the characteristics of the data set in this paper.

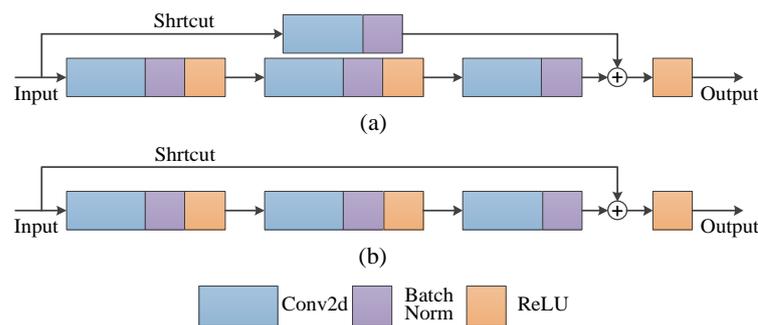


Figure 2. Residual block of ResNet50. (a) first layer structure of residual body. (b) subsequent structure of residual body.

Due to the fact that recognizing tomato young fruits obtained under different conditions is not high, and the relevant fine-grained features are difficult to capture, enhancing the effective attention of the network to the fine-grained features is the key to solving this problem. The attention mechanism assigns high-contribution information to larger weights while suppressing other irrelevant information through weight distribution, which is an effective method to improve the performance of feature extraction networks [29,30]. Based on ResNet50, this paper uses the CBAM attention module to further optimize acquired features. As shown in Figure 3, the shape of the input feature matrix is $W \times H \times C$. After Max pooling and Average pooling, 2 groups of $1 \times 1 \times C$ feature matrices are obtained, and they are passed through Multilayer Perceptron (MLP), then the 2 output feature matrices are added to obtain the weight information Channel Attention (CA) of different channels. The calculation method is shown in Equation (1). After the CA is multiplied by the input feature matrix, the feature matrix that integrated with the channel attention is obtained, as shown in Feature X' in Figure 3. The feature matrix with channel attention fused is then passed through the Max pool and Average pool of $W \times H \times 1$, respectively, to obtain 2 feature maps, and concat the 2 feature maps in depth direction, then the convolution operation is performed to obtain the Spatial Attention (SA) that integrates the spatial weight information, and the calculation method is shown in Equation (2). In Equation (2), $f^{7 \times 7}$ represents that the size of the pooling kernel is 7×7 . Finally, SA is multiplied by feature X' to obtain the feature map Refined Feature X'' , which combines channel and spatial attention information.

$$CA(X) = \sigma(\text{MLP}(\text{Max Pool}(X)) + \text{MLP}(\text{Avg Pool}(X))), \tag{1}$$

$$SA(X) = \sigma(f^{7 \times 7}([\text{Max Pool}(X'); \text{Avg Pool}(X')]), \tag{2}$$

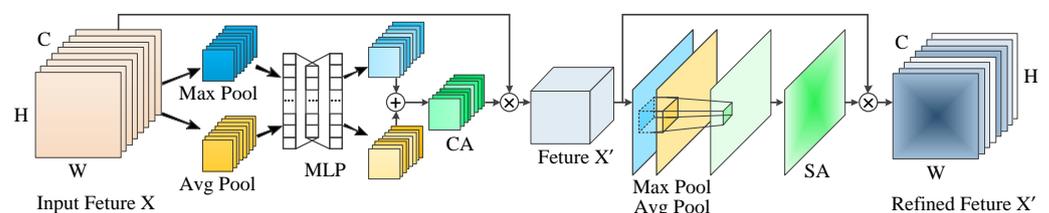


Figure 3. Convolutional Block Attention Module (CBAM).

2.3. Multi-Scale Feature Fusion

Since there may be multiple different-sized tomatoes in the image, different features are needed to distinguish objects of different scales. However, with the abstraction of features extracted by CNN, the size of its feature maps gradually shrinks, resulting in the loss of detailed information. In order to obtain robust high-level semantic and low-level detailed information at the same time, and to improve the model’s adaptation to different scale features based on different scale feature maps generated by the feature extraction

network and CBAM attention mechanism, this paper constructs a pyramid structure with strong semantic features on all scales. Through the bottom-up feature extraction and top-down feature fusion process, the semantic information and detailed information in the features are optimized to form a feature pyramid structure.

The detailed structure of the feature pyramid constructed in this paper is shown in Figure 4. The feature extraction network ResNet50 contains 4 residual bodies: C2, C3, C4 and C5. Convolutioning the original image, the last residual structure of each residual body is selected to output the feature map; therefore, a total of 4 feature maps are obtained. It should be noted that the size of the output feature map of each residual body is half of the previous residual body, and the depth of the feature matrix is doubled. The obtained feature maps of different scales are respectively optimized by the CBAM module and then input into the FPN structure. In order to ensure the normal fusion of subsequent feature maps, first, adjust the channel of the feature maps which input FPN through 1×1 convolution to ensure that the feature matrices of different scales have the same depth. Then, the top-down process in FPN performs up-sampling of the abstract semantic features twice, and fuses it with the corresponding horizontally connected feature maps of the next layer. Since this method only adds a horizontal connection to the initial network, it generates very little additional computational cost. Finally, a 3×3 convolution operation is used for each layer of fused feature maps to reduce the aliasing effect caused by up-sampling. After the above operations, the features are optimized and feature fusion is performed through the top-down method, so that feature maps of all scales have rich semantic information. The feature maps P2–P5 output through the FPN structure in Figure 4 correspond to C2–C5 in ResNet50, and P6 is a higher-level abstract feature obtained by Max pooling on the basis of P5.

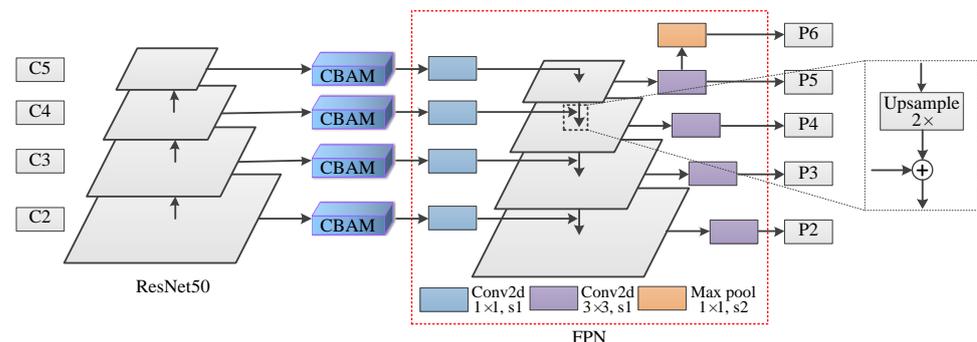


Figure 4. The structure of Feature Pyramid Network (FPN).

2.4. The Network Architecture of Improved Faster R-CNN with Attention Mechanism

The overall structure of the improved Faster R-CNN is shown in Figure 5. It is mainly composed of 4 parts, including optimized backbone, region proposal network (RPN), bounding box regression and classification. Since all of the above modules are implemented by CNN, all can run on the GPU, the detection speed is fast and the comprehensive performance of the model is better. First, the input image is standardized and scaled to a fixed size. ResNet50 with CBAM is used to extract image features and the transfer learning strategy is used to fine-tune network parameters. Then, through FPN, 4 fusion feature maps of P2, P3, P4 and P5 are obtained, which are shared in RPN and subsequent bounding box regression and classification. The P6 with a higher degree of abstraction is only used for RPN training. The RPN is used to generate region proposals, determine whether the anchors are foreground or background through softmax function and then use bounding box regression to correct the anchors to obtain an accurate bounding box. The position of the original Faster R-CNN bounding box is obtained from model regression, while the pooled feature map size required to be fixed; therefore, the region of interest pooling (RoI Pooling) operation has 2 quantization processes: one is to quantize the boundary of the bounding box into integer coordinate values, and the other is to divide the quantized boundary area into $k \times k$ units and quantize the boundaries of each unit. There is a certain

deviation between RoI after the above operation and the initial RoI, which affects the detection accuracy of small targets. In order to further improve the positioning accuracy of the bounding box, the RoI Align feature mapping method is used to improve the extraction accuracy of the RoI. Finally, the proposals are sent to the subsequent fully connected layer for classification and bounding box regression.

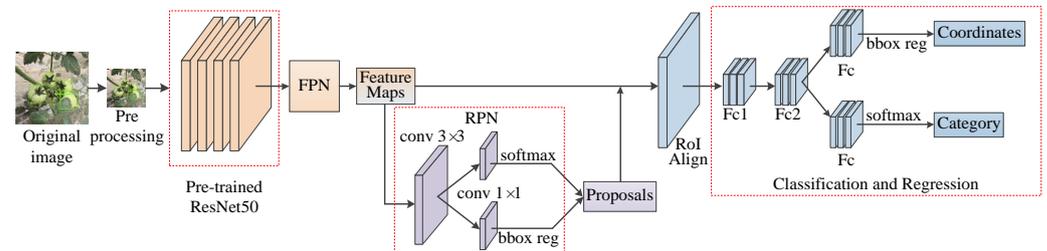


Figure 5. The overall structure of improved Faster R-CNN with Attention Mechanism.

3. Results and Discussion

3.1. Model Training Details

Experiments in this paper are trained and tested on a server equipped with GPU. The experiment equipment configuration information is shown in Table 2. The detection model is built based on the Pytorch 1.7.1 deep learning framework. The pre-training model is used to initialize the network parameters. In order to accelerate the model convergence, set a larger initial learning rate of 0.001, the weight decay rate of 0.0005, and a total of 50 epochs are trained. When the model loss no longer decreases, stop the training and save the model weight information.

Table 2. Hardware and software environment.

Configuration Item	Value
CPU	Intel® Xeon(R) Gold 5217 CPU@3.00 GHz
GPU	NVIDIA TESLA V100 (32 GB)
Operating System	Ubuntu 18.04.5 LTS 64
RAM	251.4 GB
Hard Disk	8 TB

3.2. Evaluation Indicators of Model Performance

The Precision and Recall of the model are calculated by True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN). The specific calculation methods are shown in Equations (3) and (4). Average Precision (AP) is a standard metric that measures the sensitivity of the network to the target, and it is an indicator that reflects the overall performance of the network. In Equation (5), the value ranges of Precision (P) and Recall (R) are both [0, 1]. This paper uses mean Average Precision (mAP) as the evaluation indicator for tomato young fruits detection and the calculation method is shown in Equation (6). Averaging the AP values of n different types of objects, and due to the fact that there is only one object type of tomato young fruits in the test set, so $n = 1$, mAP and AP are equal. In addition, the average detection time of the model on the test set images is also an important indicator to measure the efficiency of the model, so the time cost of the model running detection task needs to be considered at the same time for evaluation [31].

$$P = TP / (TP + FP), \quad (3)$$

$$R = TP / (TP + FN), \quad (4)$$

$$AP = \int_0^1 P(R) dR, \quad (5)$$

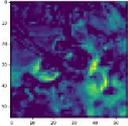
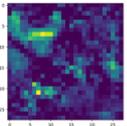
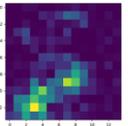
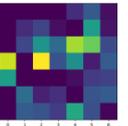
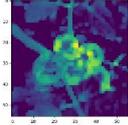
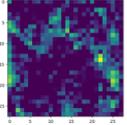
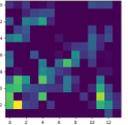
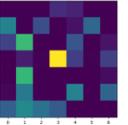
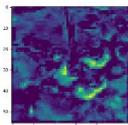
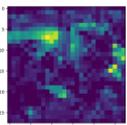
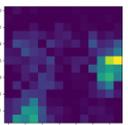
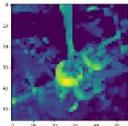
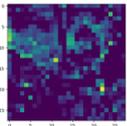
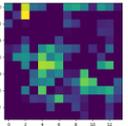
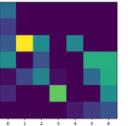
$$mAP = \frac{1}{n} \sum_n \int_0^1 P(R) dR, \tag{6}$$

3.3. Results and Discussion

3.3.1. Feature Map and Heat Map of Improved Feature Extraction Module

The iterative process of the weight information in the training process of CNN cannot be visually displayed. Therefore, in order to understand and analyze the feature extraction effects of the feature extraction network in this paper, the features extracted from the C2, C3, C4, and C5 residues body in backbone are output and visualized. The specific method is to add up the feature maps extracted by all convolution kernels in the last convolutional layer of each residual body to obtain the output feature map of the current residual body. The analysis of the feature map obtained above is helpful to understand how the continuous convolutional layer performs feature extraction and completes the conversion of the input features. Table 3 shows the output of backbone with different residuals in C2–C5, and also shows the optimization effect of the CBAM attention mechanism on the feature extraction ability of the model. Since the transfer learning fine-tune strategy is used in the model training process, it can be seen that the tomato fruits area and the background can be separated in underlying features extracted by the network. Compared with the non-attention module, backbone with CBAM has stronger feature extraction capabilities for underlying features and the position information is more accurate, which benefits from the spatial attention operation in the attention mechanism. When the depth of network increases, the features continue to be abstracted, making the high-level output feature matrix with rich abstract semantic features, but the target location is relatively rough.

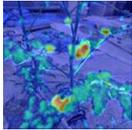
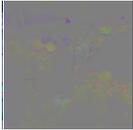
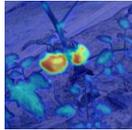
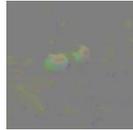
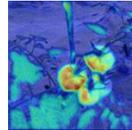
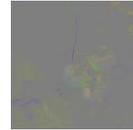
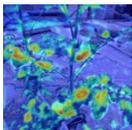
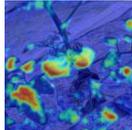
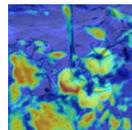
Table 3. Output feature maps of different residual body.

Original Image	CBAM	C2	C3	C4	C5
	No				
	Yes				
	No				
	Yes				

In order to verify the effectiveness of high-level semantic information extracted by the model, this paper uses Grad-CAM++ [32] to visualize the response of high-level semantic information in original images. Grad-CAM++ mainly uses the gradient information of the target to obtain the weight of the feature map, and then performs a weighted summation to obtain the heat map. The specific process is as follows: first, the image passes through the last layer of the feature extraction to obtain the feature map with the highest semantic level, and the weight of each feature map is calculated by back-propagation; then, multiply the feature map with the weight and pass the ReLU activation function to only retain useful

features; and finally, the resulting heat map is expanded to the same size with input image and weighted with it. Three test images, as shown in Table 4, are randomly selected from the test set. It can be seen from the corresponding heat map that the addition of the CBAM module improves the sensitivity of the feature extractor to high-level semantic features and can accurately focus on the fruit area. From the fine-grained feature map, it can be found that the feature extractor based on the attention mechanism retains the low-level, fine-grained information relative completely and this further confirms the effectiveness of the method.

Table 4. Heat map and Fine-Grained map of feature extractor.

Original Image						
	Heat Map	Fine-Grained	Heat Map	Fine-Grained	Heat Map	Fine-Grained
With CBAM						
Without CBAM						

3.3.2. Detection Effect of Soft-NMS on Densely Distributed Young Fruits

Non-Maximum Suppression (NMS) [33] is a method to filtrate the candidate regions generated by the region proposal network and the bounding box regression network in object detection. The specific operation method of NMS is to directly eliminate other prediction boxes that have a higher degree of coincidence with the bounding box with the highest score. However, the tomato young fruits are characterized by clusters and dense distribution and there is a large amount of fruit overlap. In this scene, the direct use of NMS will cause the object with low confidence to be missed.

In order to avoid the missed detection problem, this paper adopts a non-maximum suppression method with attenuation characteristics and considers the score and coincidence degree at the same time. Soft-NMS reduces the classification confidence scores of the overlapping bounding box in the form of weight attenuation on the basis of NMS, instead of directly removing the bounding box with higher overlap, so as to retain the bounding box with a certain degree of overlap. This method has a better detection effect for dense objects, and the calculation process is shown in Equation (7). When the iou of M and b_i is less than the given threshold N_t , S_i remains unchanged. When iou is greater than or equal to a given threshold, S_i does not take 0, but attenuates according to the linear rule. This strategy is a linear function of iou, multiplying the b_i bounding box score by a weight function, this function will attenuate the adjacent bounding box score that overlaps with the highest score bounding box M . The higher the overlap with M , the more severe the attenuation of the score. For this reason, the Gaussian function is selected as the weight function, and the rule for deleting the bounding box is modified, as shown in Equation (8), where D is the final bounding box set. Based on this method, the overlapped and occluded tomato young fruits bounding box can be retained. At the same time, because Soft-NMS is

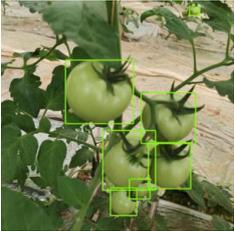
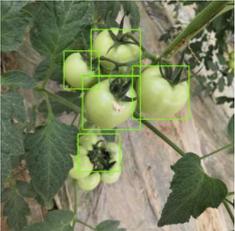
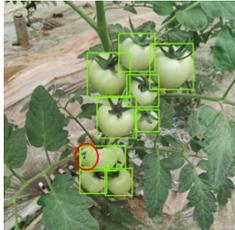
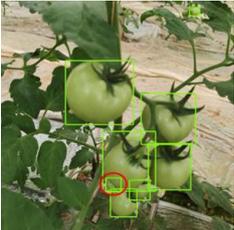
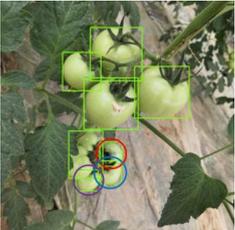
introduced into the object detection algorithm without retraining the original model, the method is very efficient.

$$S_i = \begin{cases} S_i, & \text{iou}(M, b_i) < N_t \\ S_i(1 - \text{iou}(M, b_i)), & \text{iou}(M, b_i) \geq N_t \end{cases} \quad (7)$$

$$S_i = S_i e^{-\frac{\text{iou}(M, b_i)^2}{\sigma}}, \forall b_i \text{ not } \in D, \quad (8)$$

This strategy can significantly improve the Recall of densely distributed object detection, is friendly to the selection process of different bounding boxes and does not affect the detection of small or non-overlapping targets. Randomly select images with dense distribution of fruits from the test set for testing, using NMS and Soft-NMS strategies to test, respectively. The results are shown in Table 5. In No.1, No.2 and No.3, there are fruits with a high degree of overlap at the lower left corner, and the NMS strategy directly deletes the bounding boxes of the occluded fruits, while adopting the Soft-NMS strategy, some overlapping bounding boxes are appropriately retained, as shown in the circular marking in the figure in Table 5. At the same time, the two strategies can achieve good detection results for fruits with a small degree of overlap at the top of the image, and the small objects at the upper right corner of No.2 can also be detected [33].

Table 5. Performance of different Non-Maximum Suppression algorithms.

	No.1	No.2	No.3
NMS			
Soft-NMS			

3.3.3. Improved Faster R-CNN with Attention Mechanism Detection Performance for Tomato Young Fruit

In order to verify the actual performance of the optimized model in this paper, the test set is used to verify the effect of the model. The test results are shown in Table 6. It can be seen from Table 6 that the mean Average Precision of this model for tomato young fruits reached 98.46%, which is 4.20% higher than Faster R-CNN and the Recall has been significantly improved. The average processing time for each image is only 0.084 s, which is 0.013 s less than Faster R-CNN, and it is able to meet real-time usage needs. The above results show that our method can achieve a satisfactory detection effect on tomato young fruits under near color backgrounds in the field.

Figure 6 shows the detection effect of this model on tomato young fruits in a variety of situations. The randomly selected test images included stems and leaves occlusion (Figure 6A,A'), fruits overlap (Figure 6B,B'), different light intensities (Figure 6C,C') and other complicated situations. In Figure 6A, tomato young fruits have the phenomenon of stems and leaves obscuration. Figure 6A' also includes the scene where the leaves occlude the fruits and there is an incomplete fruit at the upper right. It can be seen from Figure 6A,A'

that the model has a good recognition effect on the different occluded fruit objects and has a higher accuracy and positioning accuracy for the recognition of young fruits with near color background in the occluded state. The fruit clusters in Figure 6B,B' mainly contain the scene of fruit overlap. The use of Soft-NMS ensures that the missed detection rate of the model to identify densely distributed young fruits is low, and the generalization ability of the model is strong. Figure 6C is an image taken at noon on a sunny day, and the fruit has a higher brightness and blurry edges while Figure 6C' is an image taken in the evening. It can be seen from the test results that the model has good adaptability to fruit recognition under different lighting conditions. Otherwise, the FPN structure fuses high-level semantic features and makes the model have a good effect for small objects detection. In Figure 6C', the model is highly sensitive to changes in the target size, the small fruit targets at the upper right in the image are detected with high positioning accuracy and the fruit obscured by a large area of leaves at the upper left in the image also can be detected. From the above results, it can be seen that the method in this paper has a good detection effect to young fruits of different sizes in complex environmental conditions.

Table 6. Detection performance of the model on the test set.

Model.	Testing Time/s	Mean Average Precision/%	Mean Average Recall/%	Frames per Second	Average Testing Time/s
Faster R-CNN	22	94.26	89.64	10.22	0.097
Our method	19	98.46	94.38	11.84	0.084

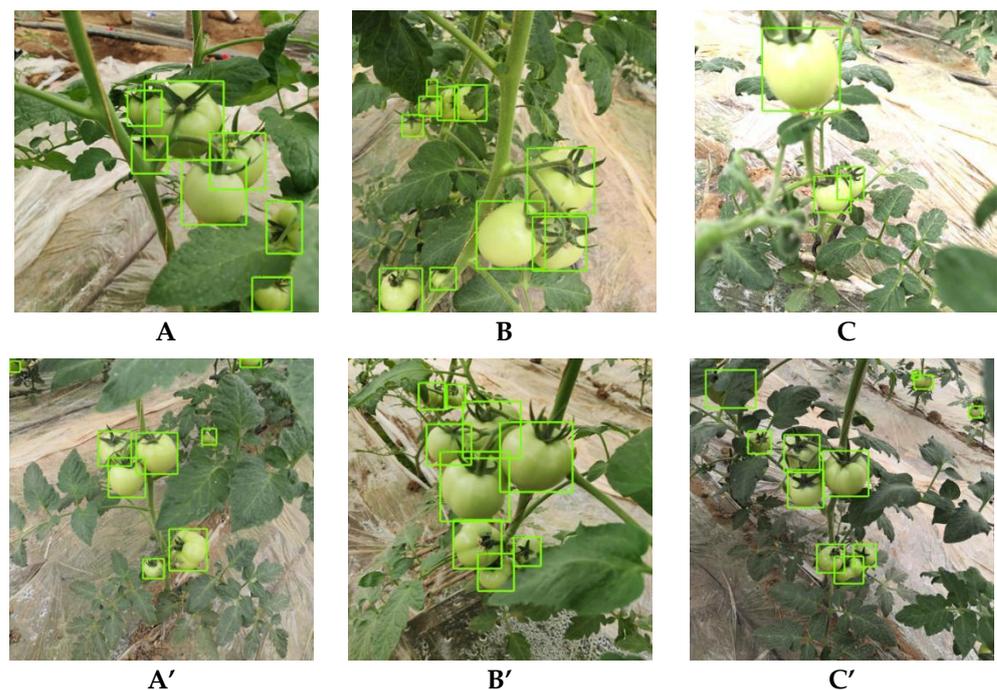


Figure 6. Detection performance of improved Faster R-CNN with Attention Mechanism under different circumstances. (A). young fruits occluded by stems and leaves; (A'). young fruits occluded by stems and leaves; (B). young fruits overlapping;(B'). young fruits overlapping; (C). image of young fruits with high light intensity; (C'). image of young fruits with low light intensity.

4. Conclusions

This paper proposes a method for detecting tomato young fruits based on Faster R-CNN. This method uses ResNet50 with a stronger feature extraction capability as backbone, accelerates model convergence based on the transfer learning fine-tune strategy and makes three improvements: (1) Introduces the CBAM attention mechanism to further optimize the output feature maps of different residuals; (2) Uses FPN to fuse high-level semantic

features with low-level detailed features to solve the problem that the model is not sensitive to small tomato young fruits and improves the scale adaptability of the model; (3) In view of the dense distribution of tomato young fruits, there are a large number of overlapping and obscured fruits, the Soft-NMS method is used to reduce the missed detection rate, so as to improve the recognition effect of clustered tomato young fruits. The test results show that the mAP of our method in the task of tomato young fruit detection reaches 98.46%, which is 4.20% higher than the original Faster R-CNN, and it can still achieve satisfactory results in a complex background. The average detection time per image in the test set is only 0.084 s, which is 0.013 s less than Faster R-CNN. The research in this paper lays a theoretical foundation for the development of automatic fruit thinning devices in orchards, and can provide references for practical applications in fruit detection and other fields.

Author Contributions: Conceptualization, P.W. and D.H.; methodology, P.W.; software, P.W.; validation, P.W. and T.N.; formal analysis, P.W.; investigation, P.W. and T.N.; resources, D.H.; data curation, D.H.; writing—original draft preparation, P.W.; writing—review and editing, P.W., T.N., and D.H.; visualization, D.H.; supervision, D.H.; project administration, D.H.; funding acquisition, D.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The Shaanxi Key Research and Development Program (CN) (Grant Number 2021ZDLNY03-02), The Key Science and Technology Program of Shaanxi Province of China (Grant Nos.: 2021NY-169).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are available on request to the corresponding author.

Acknowledgments: We all appreciate the anonymous reviewers for their valuable comments on this manuscript that allowed us to improve the quality of this paper. Also, we all appreciate Zhifeng Yao and Jin Hu for their help in the modification process.

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

References

1. Slatnar, A.; Mikulic-Petkovsek, M.; Stampar, F.; Veberic, R.; Marsic, N.K. Influence of cluster thinning on quantitative and qualitative parameters of cherry tomato. *Eur. J. Hort. Sci.* **2020**, *85*, 30–41. [[CrossRef](#)]
2. Xu, Z.; Jia, R.; Liu, Y.; Zhao, C.; Sun, H. Fast Method of Detecting Tomatoes in a Complex Scene for Picking Robots. *IEEE Access* **2020**, *8*, 55289–55299. [[CrossRef](#)]
3. Sun, J.; He, X.; Ge, X.; Wu, X.; Shen, J.; Song, Y. Detection of Key Organs in Tomato Based on Deep Migration Learning in a Complex Background. *Agriculture* **2018**, *8*, 196. [[CrossRef](#)]
4. Kang, H.; Zhou, H.; Chen, C. Visual Perception and Modeling for Autonomous Apple Harvesting. *IEEE Access* **2020**, *8*, 62151–62163. [[CrossRef](#)]
5. Li, J.; Huang, W.; Zhao, C. Machine vision technology for detecting the external defects of fruits—a review. *Imaging Sci. J.* **2015**, *63*, 241–251. [[CrossRef](#)]
6. Kamilaris, A.; Prenafeta-Boldu, F.X. Deep learning in agriculture: A survey. *Comput. Electron. Agr.* **2018**, *147*, 70–90. [[CrossRef](#)]
7. Tang, Y.; Chen, M.; Wang, C.; Luo, L.; Li, J.; Lian, G.; Zou, X. Recognition and Localization Methods for Vision-Based Fruit Picking Robots: A Review. *Front. Plant Sci.* **2020**, *11*, 510. [[CrossRef](#)] [[PubMed](#)]
8. Zhao, Y.; Gong, L.; Zhou, B.; Huang, Y.; Liu, C. Detecting tomatoes in greenhouse scenes by combining AdaBoost classifier and colour analysis. *Biosyst. Eng.* **2016**, *148*, 127–137. [[CrossRef](#)]
9. Zhao, Y.; Gong, L.; Huang, Y.; Liu, C. Robust Tomato Recognition for Robotic Harvesting Using Feature Images Fusion. *Sensors* **2016**, *16*, 173. [[CrossRef](#)]
10. Wu, J.; Zhang, B.; Zhou, J.; Xiong, Y.; Gu, B.; Yang, X. Automatic Recognition of Ripening Tomatoes by Combining Multi-Feature Fusion with a Bi-Layer Classification Strategy for Harvesting Robots. *Sensors* **2019**, *19*, 612. [[CrossRef](#)]
11. Yamamoto, K.; Guo, W.; Yoshioka, Y.; Ninomiya, S. On Plant Detection of Intact Tomato Fruits Using Image Analysis and Machine Learning Methods. *Sensors* **2014**, *14*, 12191–12206. [[CrossRef](#)]
12. Gongal, A.; Amatya, S.; Karkee, M.; Zhang, Q.; Lewis, K. Sensors and systems for fruit detection and localization: A review. *Comput. Electron. Agr.* **2015**, *116*, 8–19. [[CrossRef](#)]
13. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.

14. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [[CrossRef](#)]
15. Chao, X.; Sun, G.; Zhao, H.; Li, M.; He, D. Identification of Apple Tree Leaf Diseases Based on Deep Learning Models. *Symmetry* **2020**, *12*, 1065. [[CrossRef](#)]
16. Yang, G.; He, Y.; Yang, Y.; Xu, B. Fine-Grained Image Classification for Crop Disease Based on Attention Mechanism. *Front. Plant Sci.* **2020**, *11*, 600854. [[CrossRef](#)]
17. Quan, L.; Feng, H.; Li, Y.; Wang, Q.; Zhang, C.; Liu, J.; Yuan, Z. Maize seedling detection under different growth stages and complex field environments based on an improved Faster R-CNN. *Biosyst. Eng.* **2019**, *184*, 1–23. [[CrossRef](#)]
18. Wang, D.; Li, C.; Song, H.; Xiong, H.; Liu, C.; He, D. Deep Learning Approach for Apple Edge Detection to Remotely Monitor Apple Growth in Orchards. *IEEE Access* **2020**, *8*, 26911–26925. [[CrossRef](#)]
19. Dias, P.A.; Tabb, A.; Medeiros, H. Apple flower detection using deep convolutional networks. *Comput. Ind.* **2018**, *99*, 17–28. [[CrossRef](#)]
20. Wan, S.; Goudos, S. Faster R-CNN for multi-class fruit detection using a robotic vision system. *Comput. Netw.* **2020**, *168*, 107036. [[CrossRef](#)]
21. Afonso, M.; Fonteijn, H.; Fiorentin, F.S.; Lensink, D.; Mooij, M.; Faber, N.; Polder, G.; Wehrens, R. Tomato Fruit Detection and Counting in Greenhouses Using Deep Learning. *Front. Plant Sci.* **2020**, *11*, 571299. [[CrossRef](#)]
22. Yu, Y.; Zhang, K.; Yang, L.; Zhang, D. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. *Comput. Electron. Agr.* **2019**, *163*, 104846. [[CrossRef](#)]
23. Chen, J.; Wang, Z.; Wu, J.; Hu, Q.; Zhao, C.; Tan, C.; Teng, L.; Luo, T. An improved Yolov3 based on dual path network for cherry tomatoes detection. *J. Food Process. Eng.* **2021**, *44*, e13803. [[CrossRef](#)]
24. Wang, D.; He, D. Recognition of apple targets before fruits thinning by robot based on R-FCN deep convolution neural network. *Trans. Chin. Soc. Agric. Eng.* **2019**, *35*, 156–163.
25. Nguyen, T.T.; Vandevoorde, K.; Wouters, N.; Kayacan, E.; De Baerdemaeker, J.G.; Saeys, W. Detection of red and bicoloured apples on tree with an RGB-D camera. *Biosyst. Eng.* **2016**, *146*, 33–44. [[CrossRef](#)]
26. Lin, G.; Tang, Y.; Zou, X.; Xiong, J.; Fang, Y. Color-, depth-, and shape-based 3D fruit detection. *Precis. Agric.* **2020**, *21*, 1–17. [[CrossRef](#)]
27. Woo, S.; Park, J.; Lee, J.Y.; Kweon, S. CBAM: Convolutional Block Attention Module. *arXiv* **2018**, arXiv:1807.06521.
28. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
29. Mi, Z.; Zhang, X.; Su, J.; Han, D.; Su, B. Wheat Stripe Rust Grading by Deep Learning With Attention Mechanism and Images From Mobile Devices. *Front. Plant Sci.* **2020**, *11*, 558126. [[CrossRef](#)]
30. Wang, P.; Niu, T.; Mao, Y.; Zhang, Z.; Liu, B.; He, D. Identification of Apple Leaf Diseases by Improved Deep Convolutional Neural Networks With an Attention Mechanism. *Front. Plant Sci.* **2021**, *12*, 723294, in press. [[CrossRef](#)] [[PubMed](#)]
31. Koirala, A.; Walsh, K.B.; Wang, Z.; McCarthy, C. Deep learning—Method overview and review of use for fruit detection and yield estimation. *Comput. Electron. Agric.* **2019**, *162*, 219–234. [[CrossRef](#)]
32. Chattopadhyay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. *arXiv* **2017**, arXiv:1710.11063.
33. Neubeck, A.; Gool, L.V. Efficient Non-Maximum Suppression. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR2006), Hong Kong, China, 20–24 August 2006; pp. 850–855.