agronomy

MDPI

Article

Xiaomila Green Pepper Target Detection Method under Complex Environment Based on Improved YOLOv5s

Fenghua Wang *, Zhexing Sun, Yu Chen, Hao Zheng and Jin Jiang

Faculty of Modern Agricultural Engineering, Kunming University of Science and Technology, Kunming 650500, China; 20202214032@stu.kust.edu.cn (Z.S.); 20212214020@stu.kust.edu.cn (Y.C.); 20202214008@stu.kust.edu.cn (H.Z.); 20212214025@stu.kust.edu.cn (J.J.)
* Correspondence: 20090099@kust.edu.cn

Abstract: Real-time detection of fruit targets is a key technology of the Xiaomila green pepper (*Capsicum frutescens* L.) picking robot. The complex conditions of orchards make it difficult to achieve accurate detection. However, most of the existing deep learning network detection algorithms cannot effectively detect Xiaomila green pepper fruits covered by leaves, branches, and other fruits in natural scenes. As detailed in this paper, the Red, Green, Blue (RGB) images of Xiaomila green pepper in the green and mature stage were collected under natural light conditions for building the dataset and an improved YOLOv5s model (YOLOv5s-CFL) is proposed to improve the efficiency and adaptability of picking robots in the natural environment. First, the convolutional layer in the Cross Stage Partial (CSP) is replaced with GhostConv, the detection speed is improved through a lightweight structure, and the detection accuracy is enhanced by adding a Coordinate Attention (CA) layer and replacing Path Aggregation Network (PANet) in the neck with Bidirectional Feature Pyramid Network (BiFPN). In the experiment, the YOLOv5s-CFL model was used to detect the Xiaomila, and the detection results were analyzed and compared with those of the original YOLOv5s, YOLOv4-tiny, and YOLOv3-tiny models. With these improvements, the Mean Average Precision (mAP) of YOLOv5s-CFL is 1.1%, 6.8%, and 8.9% higher than original YOLOv5s, YOLOv4-tiny, and YOLOv3-tiny, respectively. Compared with the original YOLOv5 model, the model size is reduced from 14.4 MB to 13.8 MB, and the running speed is reduced from 15.8 to 13.9 Gflops. The experimental results indicate that the lightweight model improves the detection accuracy and has good real-time performance and application prospects in the field of picking robots.

Keywords: *Capsicum frutescens* L.; improved YOLOv5; target detection; orchard

Citation: Wang, F.; Sun, Z.; Chen, Y.; Zheng, H.; Jiang, J. Xiaomila Green Pepper Target Detection Method under Complex Environment Based on Improved YOLOv5s. *Agronomy* **2022**, *12*, 1477. https://doi.org/10.3390/agronomy12061477

Academic Editor: Roberto Marani

Received: 9 June 2022
Accepted: 16 June 2022
Published: 20 June 2022

1. Introduction

Yunnan Province is one of the three main producing areas of chili peppers in China. As a semi-domesticated small-fruited chili pepper variety, Xiaomila green pepper is mainly distributed in Honghe, Wenshan, and other places in Yunnan. The total output value of spicy food has reached CNY two billion [1,2], but the research on mechanized harvesting is still in its infancy for crops such as millet, which has the characteristics of simultaneous and batch harvesting of flowers and fruits. The harvesting process of Xiaomila green pepper is mostly performed by one household manually, which is labor-intensive and has a low production efficiency. With the reduction of the rural population and the increase of labor costs, the timely harvest of millet Xiaomila has been seriously affected, and the development of its industry has been restricted.

With the continuous development of agricultural automation technology, agricultural picking robots have shifted from research and development to the experimental stage, providing a new approach for the mechanized picking of Xiaomila. Rapid and accurate positioning and identification of ripe fruit is the focus and hot issue of picking robot research. Because the Xiaomila green pepper fruits of millet are in the shape of short cones,

short fingers, or rice grains, the peels of green and ripe fruits are light yellow-green, the peels are smooth or slightly wrinkled, and the fruit-bearing rate per plant is high and the spatial distribution is irregular, it is difficult to identify fruits in complex field environments. There are problems of different target scales, low chromatic aberration, and high occlusion in picking, which increases the perceptual judgment and picking difficulty of the machine picking system.

Currently, the research on machine picking of pepper fruit is still in its infancy. Kitamura and Oka et al. [3] identify green peppers in a greenhouse with LED light reflection. Green peppers are identified by intensity, saturation, and chromaticity thresholds according to different degrees of light reflection on the fruit and leaf surfaces. However, its applicability is limited, and the effect is only obvious in the case of weak light. Bac et al. [4] construct a recognition method for green pepper. The plants of green peppers are divided into two parts: hard barriers (stems and fruits) and soft barriers (leaves and petioles). Then, a hyperspectral camera is used to obtain plant features. Due to the changes in natural light, the light incident angles between the plants are different, and the detection rate between scenes is only 59.2%. Ji et al. [5] proposed a new algorithm based on support vector machine (SVM) to identify green peppers, and the mutation strategy was introduced to improve the particle swarm optimization algorithm. The model obtains a recognition accuracy of 89.04%, but it is difficult to find green peppers with dense growth and high occlusion. McCool et al. [6] classified sweet pepper hyperspectral image data using Conditional Random Fields (CRF). The method combines the texture features of sweet peppers including Histogram of Oriented Gradients (HOG), Local Binary Pattern (LBP), and Sparse Auto-Encoder (SAE) features, and these features are input into the CRF for training to detect target fruits. The recognition accuracy of this method is only 69.2% on real farms. Ji et al. [7] proposed a method based on manifold sorting for target recognition. The super-pixels are extracted by energy-driven sampling (SEEDS) to construct the super-pixel block of the enhanced image. Then, the image boundaries are sorted by manifold sorting, and the final saliency map is obtained by fusion. Li et al. [8] introduced a method that combines an adaptive spatial feature pyramid with an attention mechanism and proposed the idea of multi-scale prediction to improve the recognition effect of occluded and small target green peppers, with an accuracy of 96.91% and a recall of 93.85%. The above research focuses on the target recognition of pepper fruit in the greenhouse environment, which requires high ambient light. The research on green pepper recognition did not use the same dataset, and the performance of the proposed approaches in the previous experiments were difficult to compare. Furthermore, the Xiaomila green pepper fruit grows in an unstructured environment. The occluded contour features put forward higher requirements for target recognition of the Xiaomila green pepper fruit.

In the past few decades, machine vision research on the detection of fruits, medicinal materials, and vegetables has progressed rapidly. Before the introduction of deep learning theory, fruit, medicinal material, and vegetable detection methods were mostly based on traditional machine learning algorithms, (color [9,10], shape [11,12], texture [13,14] or fusion features [15,16], Support Vector Machines [17], etc.). However, these methods often lack generality and robustness.

With the continuous development of deep learning technology and the rapid improvement of GPU performance, more and more scholars consider applying lightweight deep convolutional networks to crop identification in complex environments [18,19]. Tian et al. [20] proposed an improved YOLOv3 model for detecting apples at different growth stages to fit the complex environment of orchards. The results show that the new model is better than the original YOLOv3 model and the region-based fast convolutional neural network (Faster R-CNN) model using VGG16. Wang et al. [21] developed an accurate apple fruit detection method with a small model size based on the YOLOv5s deep learning algorithm with channel pruning. Parvathi et al. [22] proposed an improved Faster R-CNN model to detect two important ripening stages of coconuts in complex backgrounds. Magalhaes et al. [23] compared the performance of five types of Single Shot MultiBox De-

tector (SSD) and You Only Look Once (YOLO) suitable for picking robots. The results show that the performance of SSD Inception v2 is the best, while the response time of YOLOv4-tiny is only 5 ms. A large number of studies have shown that deep learning technology can be used to compare with background research on the recognition method of target fruits with a similar color [24–28].

To adapt to the influence of road conditions on the quality of real-time photos taken by the picking robot during traveling, this paper designs an improved YOLOv5s model to address the problems of missed detection, occlusion, and similar colors of fruits in the natural environment. The efficient and fast target detection system of Xiaomila green pepper fruit is of great significance to realizing the automatic operation of Xiaomila green pepper picking.

2. Materials and Methods

2.1. Xiaomila Green Pepper Image Collection

2.1.1. Methods and Image Collection

This study takes the Xiaomila green pepper in the planting base in Shupi Township (104°6′44″ N, 23°53′7″ E), Qiubei, Wenshan, Yunnan Province as the research object. The RGB images of Xiaomila green pepper in the green and mature stage were collected under natural light conditions. The average plant height of millet pepper plants was 89 cm, and the average plant spacing was 49.8 cm. The farming mode of one row and three rows was adopted, which is suitable for picking robots to work in the field. The Intel RealSense D435i camera was used to capture JPEG images with an image resolution of 1920 × 1080 pixels, and the image acquisition method is shown in Figure 1. The images of Xiaomila green pepper were collected under different light conditions in the morning and afternoon, and a total of 1200 photos of Xiaomila green pepper fruits were collected.

Figure 1. Methods of image acquisition.

2.1.2. Image Preprocessing

Firstly, 840 images were randomly selected from the amplified as the training set, 240 images as the test set, and 120 images as the verification set at the ratio of 7:2:1 to perform parameter verification and deep network training to avoid overfitting of the training model. The images of Xiaomila green pepper under different conditions are shown in Figure 2.

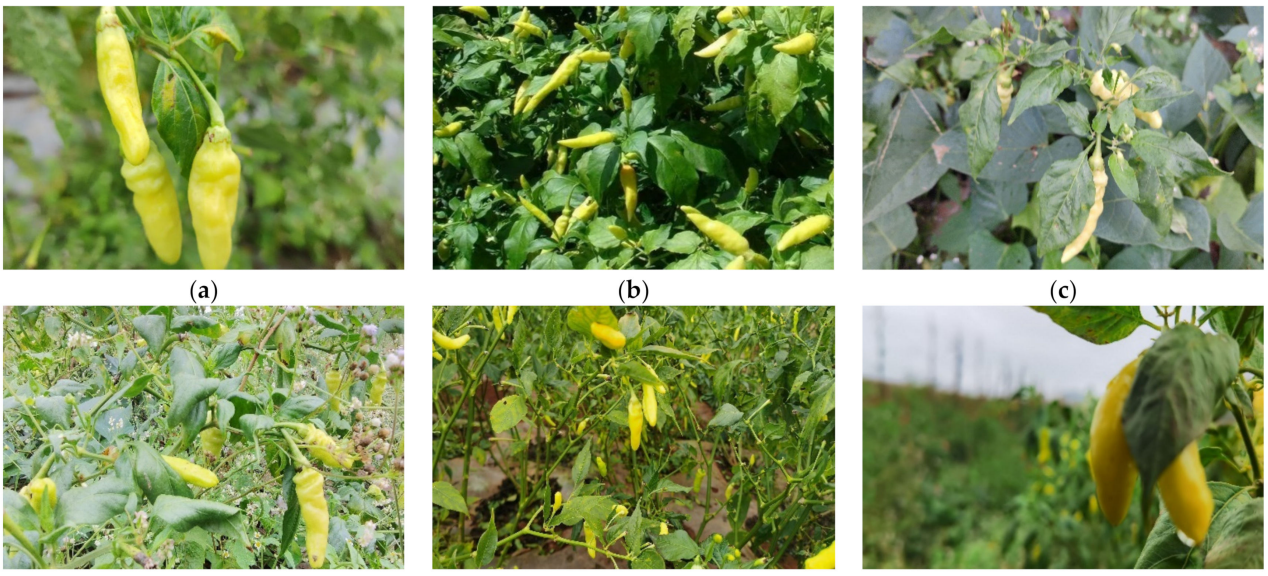


Figure 2. Images of Xiaomila green pepper under different conditions. (**a**) A single cluster of Xiaomila green pepper, cloudy, with no obvious occlusion; (**b**) a cluster of Xiaomila green peppers, sunny, with backlight; (**c**) Xiaomila green peppers covered by leaves; (**d**) in the afternoon with sufficient light, covered by branches and pedicels; (**e**) small target of Xiaomila green peppers, in the early morning, covered by branches; (**f**) fruit hanging with dew.

To perform target detection training based on deep learning with a large amount of image data, the dataset was enhanced to fit the training requirement, which can better extract image features and avoid overfitting.

Considering the impact of the complex environment of the Xiaomila green pepper picking process on fruit recognition, image rotation, image mirror flipping, image noise addition, and image brightness and contrast adjustment were performed to reduce the impact of the complex posture of the pepper fruit on the network performance. By changing the brightness and contrast of the image, the brightness deviation caused by ambient lighting changes and sensor differences was reduced, and the Cutout method was adopted to randomly select multiple fixed-size square areas to fill zero-pixel values. Meanwhile, center normalization operations were performed to simulate complex environments and remove the occlusion of the leaves on the Xiaomila green pepper fruit. The result of the data augmentation method is shown in Figure 3. The final training set consists of 8400 images, which were used as final training set data of the object recognition, including 7560 augmented images and 840 original images. There is no overlap between training and test set. The images were manually annotated with LabelImg (https://github.com/tzutalin/LabelImg (accessed on 8 June 2022)), and the smallest closed rectangle of the Xiaomila green pepper fruit is annotated (fruit with a relatively small pixel and a visible part of less than 20% were not labeled). All annotation files were saved and converted to TXT files.

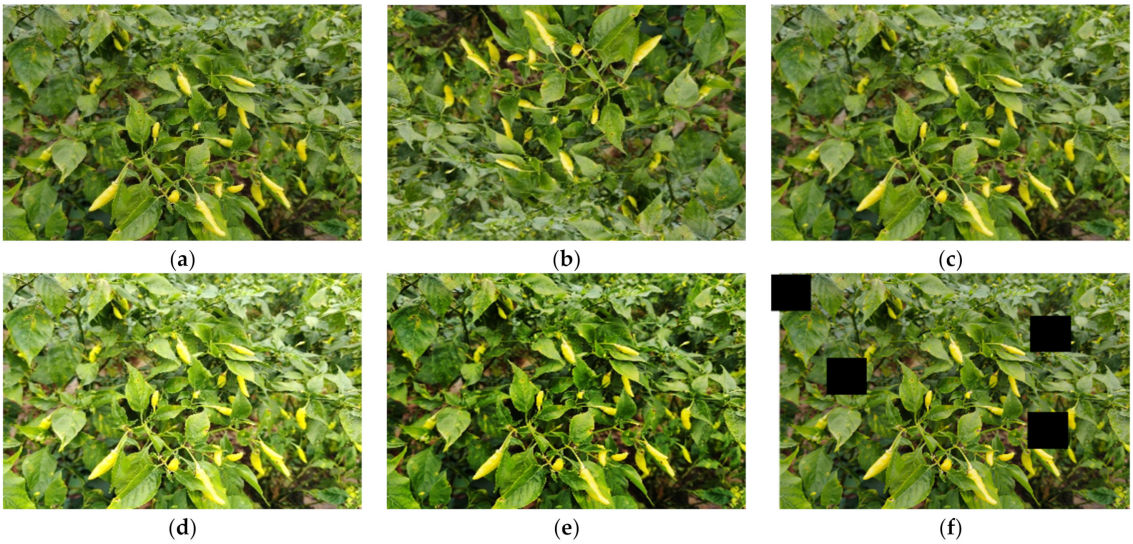


Figure 3. Image enhancement results. (**a**) Ripe Xiaomila green pepper; (**b**) rotate the image; (**c**) add noise; (**d**) improve image brightness; (**e**) improve image contrast; (**f**) cutout.

2.2. Improvements to the YOLOv5s Network Model

2.2.1. YOLOv5 Model

Previous studies [29] have demonstrated that the YOLOv5 model has outstanding performance in crop fruit recognition. The model can quickly regress image information with high detection accuracy, small model weight files, and fast training speed. It contains four architectures, YOLOv5x, YOLOv5l, YOLOv5m, and YOLOv5s, and the architecture size varies with the convolution kernel size and feature extraction time. The accuracy and real-time performance of the Xiaomila green pepper detection model are the keys to ensuring the operational efficiency of the picking robot.

The YOLOv5s framework consists of a backbone, neck, and head. The backbone forms a convolutional neural network for image feature extraction by aggregating different types of image information. The neck transfers the output image of the backbone layer in the pyramid hybrid structure to the prediction layer. The head generates prediction boxes and categories according to the image features transmitted by the neck. The basic framework of YOLOv5s is shown in Figure 4.

2.2.2. Improved Methods

The target detection algorithm of the Xiaomila green pepper picking robot has to accurately identify the Xiaomila green pepper fruit in a complex environment and reduce the model size by optimizing the YOLOv5s backbone so that it can be easily installed in the picking robot. This study aims to improve the network structure to increase the accuracy of object detection and improve the detection speed and reduce the network parameters.

The original YOLOv5s model utilizes CSP to increase the network depth and improve the network's characteristics and detection capabilities. However, in the process of testing the Xiaomila green pepper fruit of millet in the natural environment, it was found that several lightweight models obtain satisfactory test results while reducing the number of model parameters. As shown in Figure 5, to improve the network detection speed and reduce the model scale, GhostConv [30] was used in the original network to replace the Conv layer of the CSP in the backbone and neck; the modified module is named GHOST. As the basic building block of GhostConv, the CBL module is composed of Conv (convolution), BN (batch normalization), and SiLU. The core idea of GhostConv is to use low-cost convolution operations to perform conventional convolution operations on feature maps to generate basic features. Then, the deep convolutional network is used to generate

more features and combine them with the basic features to generate a large number of feature maps with the feature information of the Xiaomila green pepper.

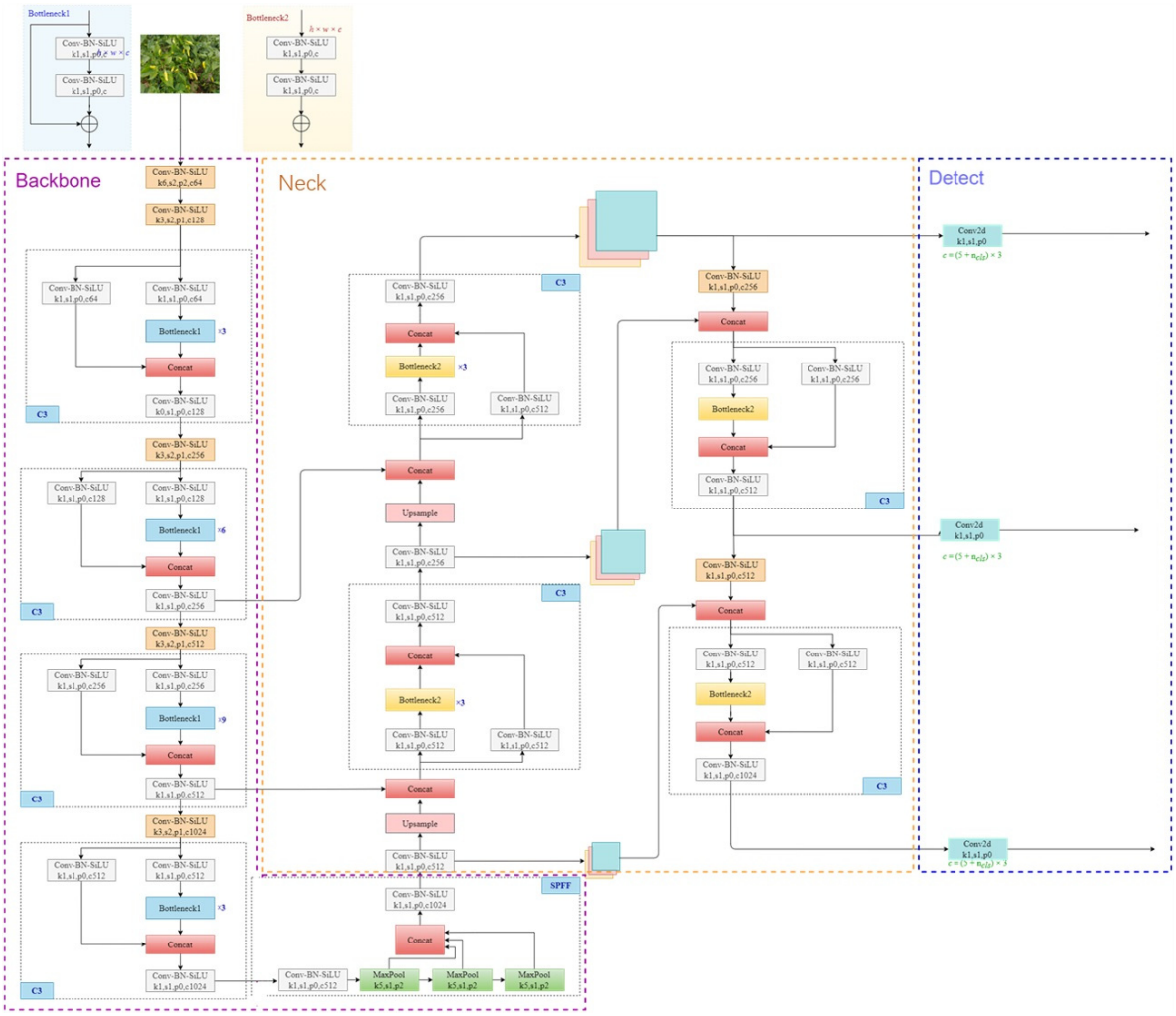


Figure 4. Framework of YOLOv5s.

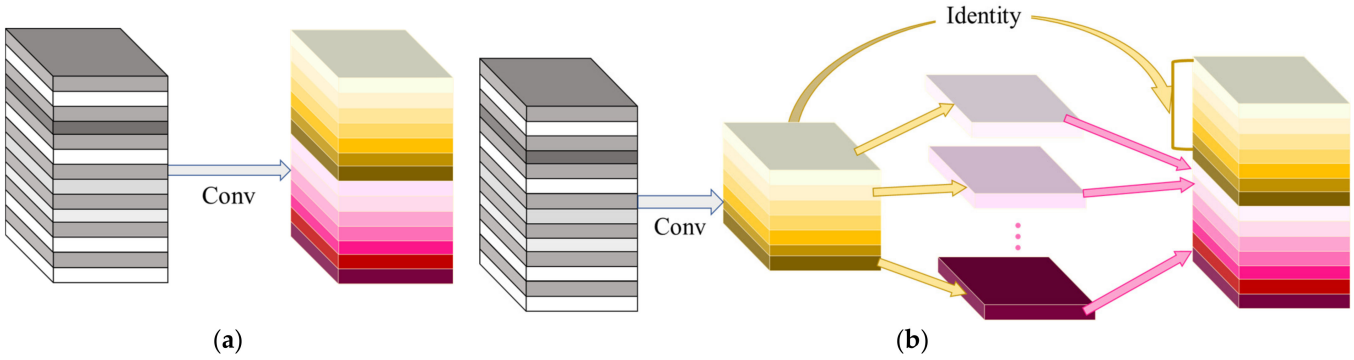


Figure 5. The feature extraction process of the model. (**a**) The convolutional layer. (**b**) The Ghost module.

Since it is difficult for current Convolutional Neural Networks (CNNs) to take features from global features, channel attention (e.g., Squeeze Excitation (SE) attention) [31] has a significant effect on improving model performance. However, this method usually ignores location information, which is important for generating spatially selective attention maps. CA [32] decomposes channel attention into two one-dimensional feature encoding processes, which aggregate features along two spatial directions respectively.

In this approach, precise location information is preserved along one spatial direction, while long-range dependencies are captured along the other spatial direction. Then, the generated feature maps are encoded as a pair of direction-aware and position-sensitive attention maps that can be applied complementarily. CA uses input feature maps to enhance the representations of objects of interest, so it performs well in image classification tasks. Considering the complexity of picking Xiaomila green pepper fruit in the orchard, as shown in Figure 6, this paper adds a CA attention mechanism layer at the end of the backbone.

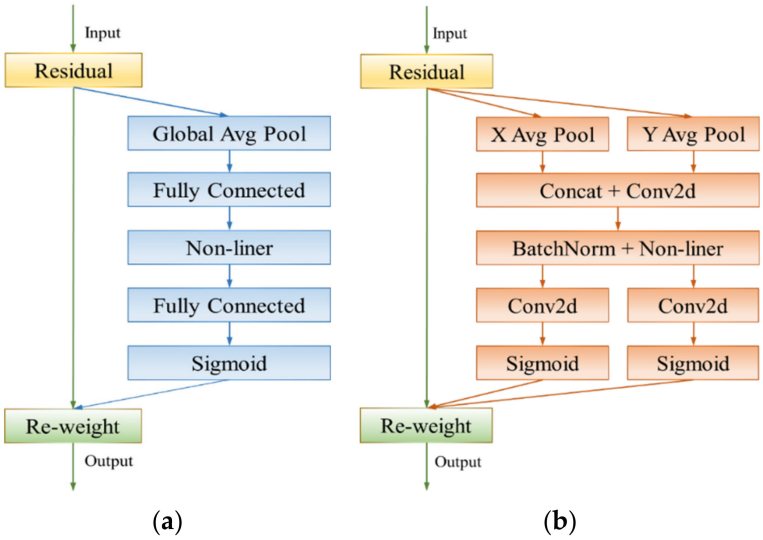


Figure 6. Comparison of the proposed coordinate attention block. (**a**) SE block. (**b**) CA block.

In general, due to the complex working environment of the picking robot, it is difficult for the acquired images to have the same initial resolution. Therefore, to identify the multi-scale Xiaomila green pepper fruit of millet, this paper changes the neck module to improve fruit detection accuracy. Although PANet in YOLOv5 achieves good results in multi-scale fusion by down-sampling and up-sampling [33], it is computationally expensive. By contrast, BiFPN can achieve a fast and simple multi-scale feature fusion. It adopts cross-scale connections to remove nodes in PANet that contribute less to feature fusion and adds additional connections between the input and output nodes at the same level [34]. This study adopts the single-layer structure BiFPN instead PANet to improve the training efficiency, as shown in Figure 7.

Since the original YOLOv5s cannot fully meet the testing requirements due to the fact that Xiaomila green pepper fruit has an irregular edge contour and the posture changes significantly. A YOLOv5s-CFL (https://github.com/01XiaoMao/CFL (accessed on 8 June 2022)) (YOLOv5s-*Capsicum frutescens* L.) model was established in this paper to detect Xiaomila green pepper fruits in complex environments. Firstly, the SiLU activation function was adopted to fit for the feature extraction of the Xiaomila green pepper, and the CA mechanism layer was added at the end of the backbone to maintain the model's feature extraction ability for deep features. In the neck, PANet was replaced with BiFPN to enhance the ability to fuse multi-scale information. Furthermore, to reduce the parameter volume and the number of network weights under the premise of ensuring the detection accuracy, the convolution layer in the CSP was replaced with GHOST to improve the detection speed while ensuring the detection accuracy and lightening the network. Its overall structure is shown in Figure 8.

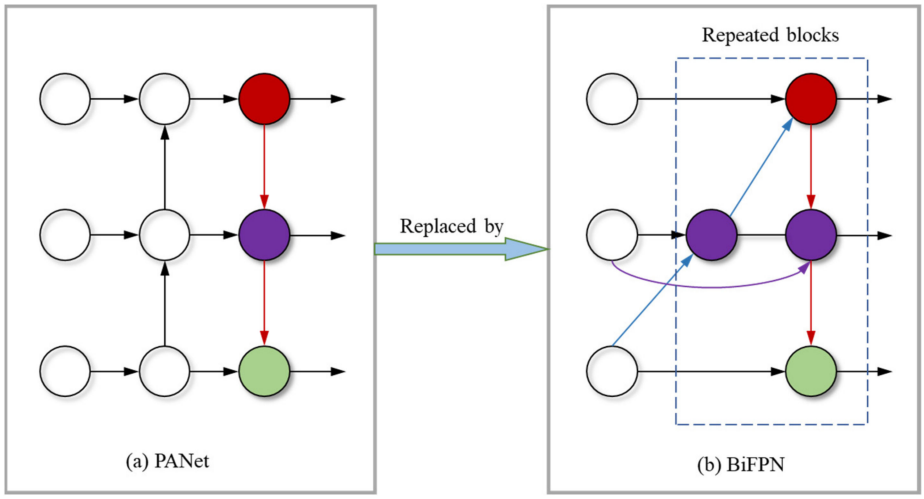


Figure 7. Using BiFPN instead of PANet to improve the feature fusion network. The colored circles represent features at different scales. (**a**) PANet adopts bottom-up and top-down paths to fuse multi-scale features; (**b**) BiFPN uses the strategy of bottom-up and top-down bidirectional feature fusion.

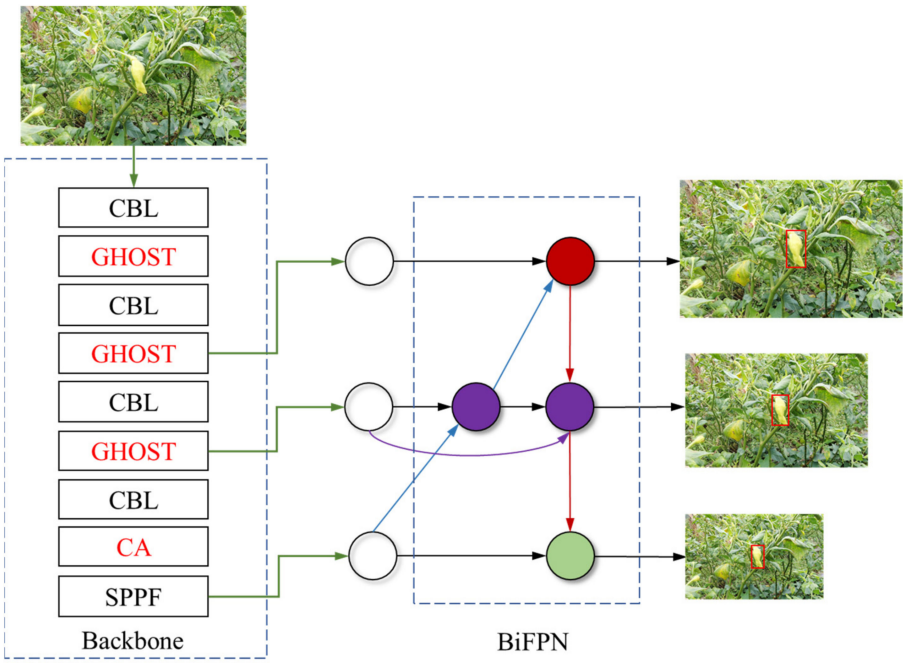


Figure 8. YOLOv5s-CFL network architecture. Modules marked in red represent improved parts in Backbone. The CSP module is replaced with GHOST. CA is inserted in the back of the backbone. The PANet is replaced with BiFPN to fuse the multi-scale features, and it is repeated once for efficiency. The outputs are on different scales.

3. Results

3.1. Training Platform

In this experiment, the Pytorch deep learning framework was built on a hardware platform equipped with Intel Xeon® W-2145 (16 GB memory, Intel Corporation, Santa Clara, CA, USA) and NVIDIA GeForce RTX2080Ti (11 GB video memory) and running Windows 10 operating system. CUDA10.2, OpenCV, Cudnn, and other related libraries were used to implement the target detection model of the Xiaomi green pepper fruit, and then the training and testing of the model were conducted.

In this study, the batch size was set to 16, and the weights of the model were regularized and updated by BN layers. The momentum was set to 0.937, and the weight decay rate

(decay rat) was set to 0.0005. The initial learning rate was set to 0.01 and IoU training threshold was set to 0.2. The training epoch was set to 450, and the relevant information was recorded after each epoch. After training, the weight file of the target detection training model was saved, and the performance of the model was evaluated on the test set. The final output of the network is the prediction candidate box for detecting the Xiaomila green pepper fruit.

3.2. Training Results

This training was iterated 450 times in total, and its loss change curve consists of two parts: bounding box loss (L_{CIoU}) and confidence loss (L_{conf}).

Because our method makes changes to the model structure, the official pre-trained weights for YOLOv5 cannot be used. Therefore, the improved YOLOv5s model was trained without pre-weighting. Meanwhile, the training data were saved and updated to the latest weight file to pre-train the weights, and the training was resumed after a training interruption. The training data of each iteration were saved to compare and analyze the performance of each model.

To explore the influence of different improvement methods on the model detection accuracy, different combinations of models were tested, and the test results are presented in Figure 9.

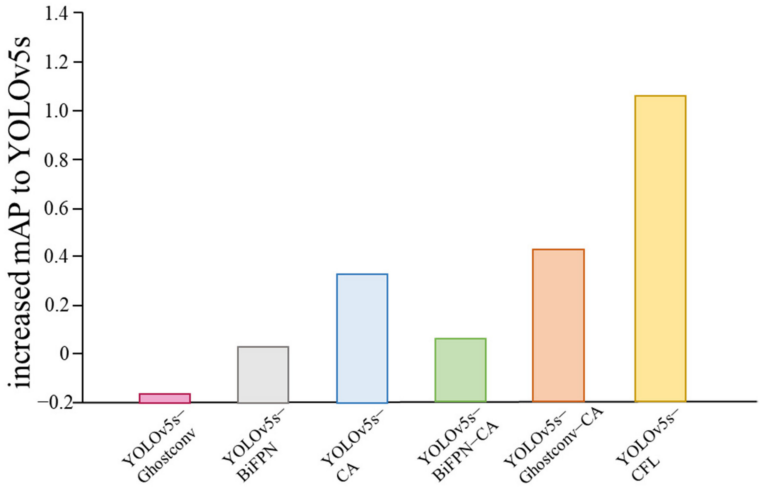
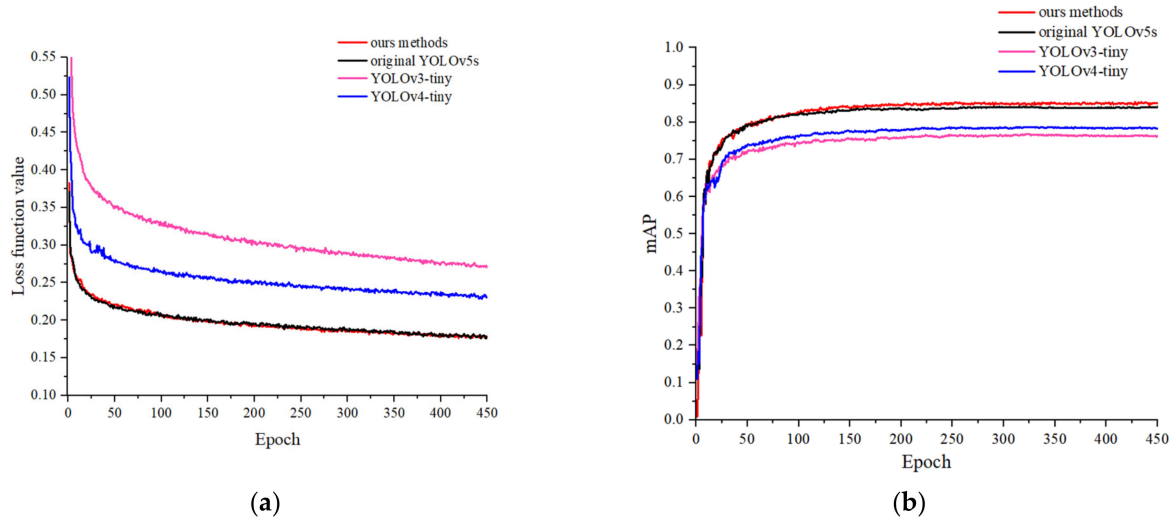


Figure 9. Increased mAP compared to the original YOLOv5s model by each combination using Xiaomila green pepper dataset.

The above results and analysis indicate that the improved YOLOv5s-CFL model can improve detection accuracy. This paper further discusses the detection results and compares them with those of the YOLO model that is widely used in other agricultural fields. Based on this, the most suitable network for the detection of Xiaomila green pepper fruit is determined.

Compared with the original YOLOv5s, the improved model shortens the training time, and the training loss curves of the four detection models are illustrated in Figure 10. Compared with the YOLOv4-tiny and YOLOv3-tiny models, the YOLOv5s and YOLOv5s-CFL models obtain lower losses. After 150 epochs, the four models gradually stabilized. The variation trend of the convergence curve indicates that the model can learn the target features of the Xiaomila green pepper fruit well, and the loss value is small after stabilization.



(**a**) (**b**)

Figure 10. The training results of the four models are compared. (**a**) Loss curve. (**b**) mAP curve.

In this paper, the model performance was evaluated by mean average precision (mAP) and F1 value. The F1 value comprehensively considers the accuracy and recall, which can reflect the stability of the model. The larger the value, the more stable the model. The calculation formula of the F1 value is:

$$\mathrm{mAP} = \frac{1}{\mathrm{C}} \int_0^1 \mathrm{P(R)dR} \tag{1}$$

$$\mathrm{F1} = \frac{2 \times \mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}} \tag{2}$$

P and R respectively refer to the accuracy and recall of the detection model, and the calculation formula is:

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \times 100\% \tag{3}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100\% \tag{4}$$

Among them, TP, FP, and FN are the abbreviations for true-positive, false-positive, and false-negative, respectively.

The detection of pepper fruits of YOLOv5s-CFL, YOLOv5s, YOLOv4-tiny, and YOLOv3-tiny models was evaluated on the validation set. The evaluation results are shown in Table 1. It can be seen from the table that the mAP value of the YOLOv5s-CFL model is 85.1%, which is 1.1% higher than that of YOLOv5s (84.0%), 6.8% higher than that of YOLOv4-tiny (78.3%), and 8.9% higher than that of YOLOv3-tiny (76.2%). The experimental results show that YOLOv5s-CFL achieves the best performance among the four models.

Table 1. Comparison of different detection models on the *Capsicum frutescens* L. dataset.

Model	Precision (%)	Recall (%)	F1 (%)	mAP (%)	Gflops	Model Size (MB)
YOLOv3-tiny	81.9	67.5	74.1	76.2	12.9	39.4
YOLOv4-tiny	82.5	71.6	76.6	78.3	16.4	21.8
YOLOv5s	82.9	74.1	78.3	84.0	15.8	14.4
YOLOv5s-CFL	83.7	74.6	78.9	85.1	13.9	13.8

Comparing the layers and weight file sizes of the four models, the YOLOv5s-CFL model has a size of 13.8 MB and a running speed of 13.9 Gflops. Compared with YOLOv5s

(14.4 MB, 15.8 Gflops), the model size is reduced, and the model parameters are reduced by nearly half.

Based on the above results and the overall analysis, the improvement of the YOLOv5s-CFL model can enhance the detection accuracy while reducing the training time and model size, thus realizing a lightweight detection model.

3.3. Detecting Results

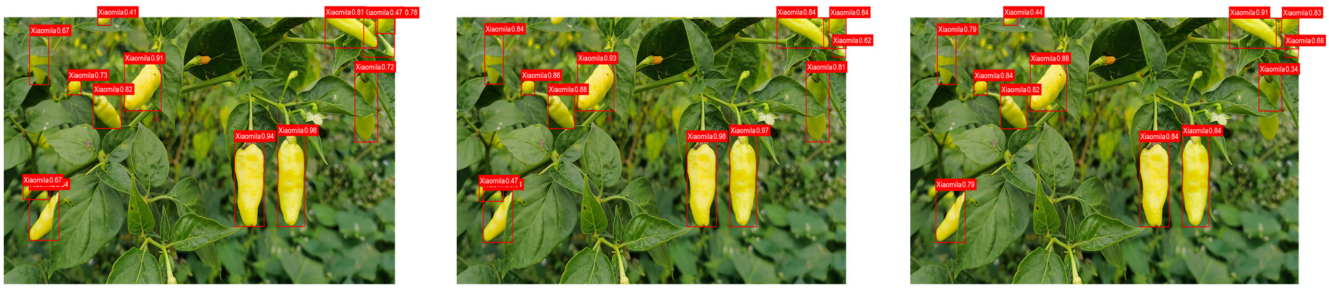
In this study, the performance of YOLOv5s-CFL, YOLOv5s, YOLOv4-tiny, and YOLOv3-tiny models for the spicy fruit of millet under complex environmental conditions was tested and analyzed.

Among the original 120 images in the test set, there are a total of 1042 Xiaomila green pepper fruits. The 44 images taken in the afternoon with sufficient light contain 438 Xiaomila green pepper fruit labels; the 56 images captured in the early morning contain 604 Xiaomila green pepper fruit labels. The YOLOv5s-CFL and YOLOv5s, YOLOv4-tiny, and YOLOv3-tiny models were applied to the detection of Xiaomila green pepper fruits in different environments, and the number of correct detections, false detections, and missing objects was counted, as shown in Table 2.

Table 2. The detection results of the four methods with different conditions.

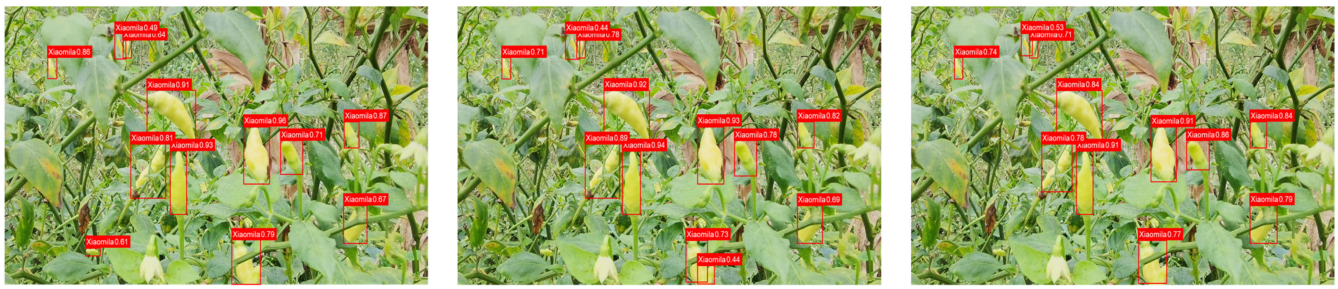
Conditions	Model	Count	Correctly Detected	Falsely Detected	Missed
Morning	YOLOv3-tiny	604	417	92	187
	YOLOv4-tiny	604	433	92	171
	YOLOv5s	604	450	93	154
	YOLOv5s-CFL	604	452	88	152
Afternoon	YOLOv3-tiny	438	286	63	152
	YOLOv4-tiny	438	313	66	125
	YOLOv5s	438	322	67	116
	YOLOv5s-CFL	438	325	63	113

In the early morning, the natural light is weak, and the lack of brightness increases the difficulty of detection. In the afternoon, the natural light is strong, the object features are easier to be captured, and most fruits can be recognized. Therefore, whether a model can detect fruit targets under different lighting conditions robustly is an important indicator to evaluate the quality of the model. As shown in Figures 11 and 12, under different light intensities, the YOLOv5s-CFL model achieves good performance in fruit detection. The detection results of YOLOv5s-CFL and YOLOv5s are relatively close and much better than those of YOLOv4-tiny and YOLOv3-tiny, and YOLOv3-tiny has the worst recognition effect. The above results show that the improved model is robust in different environments.



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

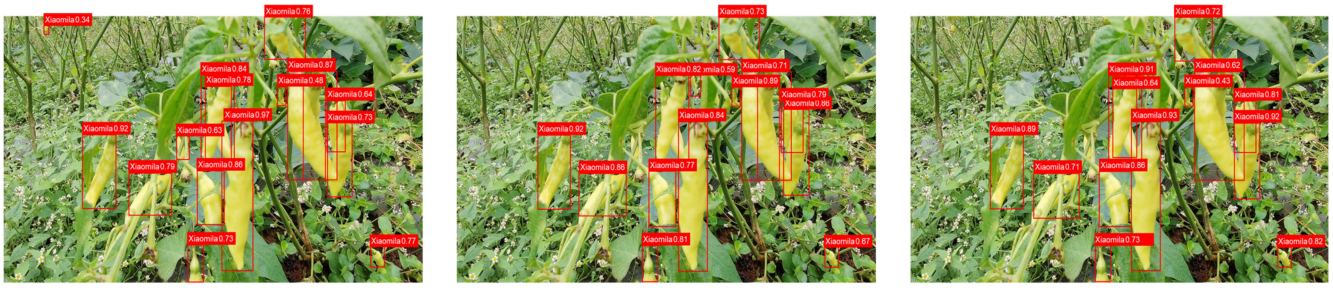
Figure 11. Comparison of the detection results of different models in the morning with YOLOv5s-CFL, YOLOv5s, and YOLOv3-tiny respectively. (**a**) YOLOv5s-CFL. (**b**) YOLOv5s. (**c**) YOLOv3-tiny.



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

Figure 12. Comparison of the detection results of different models in the afternoon with YOLOv5s-CFL, YOLOv5s, and YOLOv3-tiny respectively. (**a**) YOLOv5s-CFL. (**b**) YOLOv5s. (**c**) YOLOv3-tiny.

In multi-scale detection, although the YOLOv5s detection model has an excellent performance in scale matching, the targets in the Xiaomila green pepper dataset are densely distributed, and the targets of various sizes are often alternately distributed, as shown in Figure 13. Furthermore, when feature fusion is performed, the negative sample area of the small target detection layer may appear as a positive sample in other effective feature layers, and the conflict between the positive samples and negative samples of each effective feature layer is more obvious in the Xiaomila dataset, as shown in Figure 14.



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

Figure 13. Comparison of the detection results of different models when large and small objects are alternately distributed. (**a**) YOLOv5s-CFL detected many unlabeled small targets; (**b**) YOLOv5s detected few unlabeled small targets but missed one fruit covered by other fruit; (**c**) YOLOv3-tiny did not find small targets.

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

Figure 14. Comparison of the detection results of different models in small target detection. (**a**) YOLOv5s-CFL detected all labeled targets; (**b**) YOLOv5s missed one fruit covered by leaves; (**c**) YOLOv3-tiny missed fewer fruits covered by leaves.

Aiming at the problem of branch and leaf occlusion, misidentified samples are rare when the fruit of Xiaomila green pepper is large. When the Xiaomila green pepper fruit has a similar size to the leaves, it is difficult to detect even with the human eye because the

fruit color and size are very similar to the background leaves. The test results show that the four models have different performances. During the detection process, when the size and color of the fruit and the leaf are similar, the YOLOv3-tiny model incorrectly judges that the leaf and pedicel as the Xiaomila green pepper fruit, as shown in Figure 15. In terms of missed detection, the lightweight model achieves the highest missed detection rate for the Xiaomila fruits.

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

Figure 15. Comparison of the detection results of different models when size and color of the fruit and the leaf are similar. (**a**) YOLOv5s-CFL detected small targets but mistakenly detected pedicel as the target; (**b**) YOLOv5s missed part of small targets and mistakenly detected pedicel as the target; (**c**) YOLOv3-tiny mistakenly detected leaf and pedicel as the target.

The above analysis indicates that the YOLOv5s-CFL model reduces the model weight of YOLOv5s on the premise of ensuring detection accuracy. The YOLOv5s-CFL model achieves better performance in detecting small targets and occluded Xiaomila green pepper fruits. It can be seen from the comparative experiments that the proposed YOLOv5s-CFL has advantages in detection accuracy, detection efficiency, and detection area setting. The improvements in the model provide support for real-time positioning and detection of Xiaomila green pepper fruits.

4. Discussion

From the above experimental results, it can be seen that the YOLOv5s-CFL model reduces the model weight of YOLOv5s on the premise of improving the accuracy. Compared with the widely used YOLOv3-tiny and YOLOv4-tiny models, this model greatly improves the performance of the detection model.

We further compare the detection results of the YOLOv5s-CFL model with the detection results of other detection methods [4–6] under our dataset by using shape, size, and color (33 features), as shown in the Table 3. The correctly detected rates of YOLOv5s-CFL are higher than traditional computer vision approaches such as [4–6].

Table 3. The detection results of our methods and traditional computer vision approaches.

Conditions	Model	Count	Correctly Detected	Falsely Detected	Missed
Morning	YOLOv5s-CFL	604	452	88	152
	CART [4]	604	298	96	306
	PSO-LSSVM [5]	604	375	109	229
	CRF [6]	604	364	112	240
Afternoon	YOLOv5s-CFL	438	325	63	113
	CART [4]	438	241	98	197
	PSO-LSSVM [5]	438	305	83	133
	CRF [6]	438	298	84	140

Furthermore, we applied YOLOv5s-CFL to the dataset provided in [8], as shown in the Table 4. It can be seen from the table that the mAP value of the YOLOv5s-CFL is higher than that of improved YOLOv4-tiny [8], and the model size is smaller.

Table 4. The detection results of YOLOv5s-CFL and improved YOLOv4-tiny on green pepper dataset.

Model	Precision (%)	Recall (%)	F1 (%)	mAP (%)	Model Size (MB)
Improved YOLOv4-tiny [8]	96.91	93.85	0.95	95.11	30.9
YOLOv5s-CFL	97.52	93.73	0.96	95.46	13.8

It can be seen from the results that the model proposed in this study has achieved good results in terms of detection accuracy and detection time.

Future Work

Specific GPU for embedded systems is widely used in the field of agricultural informatization. We will also consider porting the model to embedded systems in future work. In addition, in order to adapt to the different postures of Xiaomila green pepper fruits for the real-time identification of the target, the next research should be combined with the motion control strategy of the grasping end effector, so as to realize the fruit picking that is covered by branches or other fruits by adjusting the picking angle and the position of the end effectors.

5. Conclusions

This paper proposes a method that can effectively detect and identify the Xiaomila green pepper fruits in the natural environment. This method is based on the YOLOv5s algorithm. It replaces the convolutional layer in the CSP module with GhostConv and adds a CA layer. Meanwhile, it replaces PANet with BiFPN and improves the detection speed through a lightweight network structure while ensuring network accuracy. In addition, the detection performance of several classic target detection networks is also analyzed. The experimental results indicate that the feature extraction and multi-scale detection effects of the improved model are significantly enhanced, and the number of training model parameters is reduced to improve the detection speed. Good results have been achieved on the Xiaomila green pepper fruit dataset. In future work, we will focus on the detection of the stalk of the Xiaomila green pepper fruit and combine the picking point positioning algorithm with the stalk and fruit detection algorithms to realize real-time positioning and detection of the picking point of the Xiaomila green pepper fruit.

Author Contributions: Collected data on Xiaomila green pepper fruit, Z.S., Y.C. and H.Z.; analyzed the data, Z.S., Y.C. and J.J.; wrote the paper, Z.S. and F.W.; drew pictures for this paper, Z.S., Y.C., J.J. and H.Z.; reviewed and edited the paper. Z.S., F.W., Y.C., J.J. and H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Key Research and Development Program (2017YFD0700600–2017YFD0700604), the Yunnan Major Science and Technology Special Program (2018ZC001-1, 2018ZC001-3, 2018ZC001-4, 2018ZC001-5), and The National Natural Science Foundation of China (51975265).

Data Availability Statement: The raw data required to reproduce these findings cannot be shared at this time as the data also form part of an ongoing study.

Acknowledgments: The authors would like to thank all the reviewers who participated in the review.

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

References

1. Ye, Z.; Shang, Z.X.; Li, M.Q.; Ren, H.B.; Qu, Y.H.; Hu, X.S.; Yi, J.J. Comparison and comprehensive analysis of quality characteristics of fermented Xiaomila in different cultivars. *Food Ferment. Ind.* **2021**, *47*, 87–95.
2. Elkhedir, A.E.; Iqbal, A.; Zogona, D.; Mohammed, H.H.; Murtaza, A.; Xu, X.Y. Apigenin glycosides from green pepper enhance longevity and stress resistance in Caenorhabditis elegans. *Nutr. Res.* **2022**, *102*, 23–34. [CrossRef] [PubMed]
3. Kitamura, S.; Oka, K.; Ikutomo, K.; Kimura, Y.; Taniguchi, Y. A Distinction Method for Fruit of Sweet Pepper Using Reflection of LED Light. In Proceedings of the Annual Conference of the SICE, Chofu, Japan, 20–22 August 2008; pp. 460–463.
4. Bac, C.W.; Hemming, J.; van Henten, E.J. Robust pixel-based classification of obstacles for robotic harvesting of sweet-pepper. *Comput. Electron. Agric.* **2013**, *96*, 148–162. [CrossRef]
5. Ji, W.; Chen, G.Y.; Xu, B.; Meng, X.L.; Zhao, D. Recognition Method of Green Pepper in Greenhouse Based on Least-Squares Support Vector Machine Optimized by the Improved Particle Swarm Optimization. *IEEE Access* **2019**, *7*, 119742–119754. [CrossRef]
6. McCool, C.; Sa, I.; Dayoub, F.; Lehnert, C.; Perez, T.; Uperoft, B. Visual Detection of Occluded Crop: For automated harvesting. In Proceedings of the 2016 IEEE International Conference on Robotics and Automation, Stockholm, Sweden, 16–21 May 2016; pp. 2506–2512.
7. Ji, W.; Gao, X.X.; Chen, G.Y.; Zhao, D. Target recognition method of green pepper harvesting robot based on manifold ranking. *Comput. Electron. Agric.* **2020**, *177*, 105663. [CrossRef]
8. Li, X.; Pan, J.D.; Xie, F.P.; Zeng, J.P.; Li, Q.; Huang, X.J.; Liu, D.W.; Wang, X.S. Fast and accurate green pepper detection in complex backgrounds via an improved Yolov4-tiny model. *Comput. Electron. Agric.* **2021**, *191*, 106503. [CrossRef]
9. Wang, C.; Tang, Y.; Zou, X.; Luo, L.; Chen, X. Recognition and Matching of Clustered Mature Litchi Fruits Using Binocular Charge-Coupled Device (CCD) Color Cameras. *Sensors* **2017**, *17*, 2564. [CrossRef]
10. Fu, L.; Tola, E.; Al-Mallahi, A.; Li, R.; Cui, Y. A novel image processing algorithm to separate linearly clustered kiwifruits. *Biosyst. Eng.* **2019**, *183*, 184–195. [CrossRef]
11. Zhu, Y.L.; Zhang, F.J.; Li, L.X.; Lin, Y.H.; Zhang, Z.X.; Shi, L.; Tao, H.; Qin, T. Research on Classification Model of Panax notoginseng Taproots Based on Machine Vision Feature Fusion. *Sensors* **2021**, *21*, 7945. [CrossRef]
12. Cubero, S.; Diago, M.P.; Blasco, J.; Tardáguila, J.; Millán, B.; Aleixos, N. A new method for pedicel/peduncle detection and size assessment of grapevine berries and other fruits by image analysis. *Biosyst. Eng.* **2014**, *117*, 62–72. [CrossRef]
13. Wang, C.; Lee, W.S.; Zou, X.; Choi, D.; Gan, H.; Diamond, J. Detection and counting of immature green citrus fruit based on the Local Binary Patterns (LBP) feature using illumination-normalized images. *Precis. Agric.* **2018**, *19*, 1062–1083. [CrossRef]
14. Nuske, S.; Wilshusen, K.; Achar, S.; Yoder, L.; Narasimhan, S.; Singh, S. Automated Visual Yield Estimation in Vineyards. *J. Field Robot.* **2014**, *31*, 837–860. [CrossRef]
15. 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] [PubMed]
16. Tao, Y.; Zhou, J. Automatic apple recognition based on the fusion of color and 3D feature for robotic fruit picking. *Comput. Electron. Agric.* **2017**, *142*, 388–396. [CrossRef]
17. Fu, L.; Duan, J.; Zou, X.; Lin, G.; Song, S.; Ji, B.; Yang, Z. Banana detection based on color and texture features in the natural environment. *Comput. Electron. Agric.* **2019**, *167*, 105057. [CrossRef]
18. Ren, S.Q.; He, K.M.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the 29th Annual Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015.
19. Sa, I.; Ge, Z.Y.; Dayoub, F.; Upcroft, B.; Perez, T.; McCool, C. DeepFruits: A Fruit Detection System Using Deep Neural Networks. *Sensors* **2016**, *16*, 1222. [CrossRef]
20. Tian, Y.N.; Yang, G.D.; Wang, Z.; Wang, H.; Li, E.; Liang, Z.Z. Apple detection during different growth stages in orchards using the improved YOLO-V3 model. *Comput. Electron. Agric.* **2019**, *157*, 417–426. [CrossRef]
21. Wang, D.D.; He, D.J. Channel pruned YOLO V5s-based deep learning approach for rapid and accurate apple fruitlet detection before fruit thinning. *Biosyst. Eng.* **2021**, *210*, 271–281. [CrossRef]
22. Parvathi, S.; Selvi, S.T. Detection of maturity stages of coconuts in complex background using Faster R-CNN model. *Biosyst. Eng.* **2021**, *202*, 119–132. [CrossRef]
23. Magalhaes, S.A.; Castro, L.; Moreira, G.; dos Santos, F.N.; Cunha, M.; Dias, J.; Moreira, A.P. Evaluating the Single-Shot MultiBox Detector and YOLO Deep Learning Models for the Detection of Tomatoes in a Greenhouse. *Sensors* **2021**, *21*, 3569. [CrossRef]
24. He, Z.L.; Xiong, J.T.; Lin, R.; Zou, X.J.; Tang, L.Y.; Yang, Z.G.; Liu, Z.; Song, G. A method of green litchi recognition in natural environment based on improved LDA classifier. *Comput. Electron. Agric.* **2017**, *140*, 159–167. [CrossRef]
25. Li, Y.T.; Sun, J.; Wu, X.H.; Lu, B.; Wu, M.M.; Dai, C.X. Grade Identification of Tieguanyin Tea Using Fluorescence Hyperspectra and Different Statistical Algorithms. *J. Food Sci.* **2019**, *84*, 2234–2241. [CrossRef] [PubMed]
26. Fu, L.H.; Yang, Z.; Wu, F.Y.; Zou, X.J.; Lin, J.Q.; Cao, Y.J.; Duan, J.L. YOLO-Banana: A Lightweight Neural Network for Rapid Detection of Banana Bunches and Stalks in the Natural Environment. *Agronomy* **2022**, *12*, 391. [CrossRef]
27. Gan, H.; Lee, W.S.; Alchanatis, V.; Ehsani, R.; Schueller, J.K. Immature green citrus fruit detection using color and thermal images. *Comput. Electron. Agric.* **2018**, *152*, 117–125. [CrossRef]
28. Jia, W.K.; Tian, Y.Y.; Luo, R.; Zhang, Z.H.; Lian, J.; Zheng, Y.J. Detection and segmentation of overlapped fruits based on optimized mask R-CNN application in apple harvesting robot. *Comput. Electron. Agric.* **2020**, *172*, 105380. [CrossRef]

29. Xu, Z.B.; Huang, X.P.; Huang, Y.; Sun, H.B.; Wan, F.X. A Real-Time Zanthoxylum Target Detection Method for an Intelligent Picking Robot under a Complex Background, Based on an Improved YOLOv5s Architecture. *Sensors* **2022**, *22*, 682. [CrossRef]
30. Han, K.; Wang, Y.H.; Tian, Q.; Guo, J.Y.; Xu, C.J.; Xu, C. GhostNet: More Features from Cheap Operations. In Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 1577–1586.
31. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-excitation networks. *arXiv* **2017**, arXiv:1709.01507.
32. Hou, Q.B.; Zhou, D.Q.; Feng, J.S. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 19–15 June 2021; pp. 13708–13717.
33. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
34. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787.