agronomy

MDPI

Article

Lightweight Blueberry Fruit Recognition Based on Multi-Scale and Attention Fusion NCBAM

Wenji Yang [1,*], Xinxin Ma [1], Wenchao Hu [1] and Pengjie Tang [2]

[1] School of Software, Jiangxi Agricultural University, Nanchang 330045, China
[2] School of Electronics and Information Engineering, Jinggangshan University, Ji'an 343009, China
* Correspondence: ywenji614@163.com

Abstract: Blueberries are widely planted because of their rich nutritional value. Due to the problems of dense adhesion and serious occlusion of blueberries during the growth process, the development of automatic blueberry picking has been seriously hindered. Therefore, using deep learning technology to achieve rapid and accurate positioning of blueberries in the case of dense adhesion and serious occlusion is one of the key technologies to achieve the automatic picking of blueberries. To improve the positioning accuracy, this paper designs a blueberry recognition model based on the improved YOLOv5. Firstly, the blueberry dataset is constructed. On this basis, we design a new attention module, NCBAM, to improve the ability of the backbone network to extract blueberry features. Secondly, the small target detection layer is added to improve the multi-scale recognition ability of blueberries. Finally, the C3Ghost module is introduced into the backbone network, which reduces the number of model parameters while ensuring the accuracy, thereby reducing the complexity of the model to a certain extent. In order to verify the effectiveness of the model, this paper conducts experiments on the self-made blueberry dataset, and the mAP is 83.2%, which is 2.4% higher than the original network. It proves that the proposed method is beneficial to improve the blueberry recognition accuracy of the model.

Keywords: object detection; YOLOv5; blueberry recognition; deep learning; multi-scale; attention mechanism

Citation: Yang, W.; Ma, X.; Hu, W.; Tang, P. Lightweight Blueberry Fruit Recognition Based on Multi-Scale and Attention Fusion NCBAM. *Agronomy* **2022**, *12*, 2354. https://doi.org/10.3390/agronomy12102354

Academic Editors: Chao Chen, Satoru Sakai, Yaqoob Majeed and Longsheng Fu

Received: 13 September 2022
Accepted: 26 September 2022
Published: 29 September 2022

1. Introduction

Blueberry is rich in nutrients and has high economic value. The blueberry-growing industry spreads all over the world [1], and more than 30 countries and regions are developing the blueberry industry. China's blueberry industry has developed rapidly over the past 20 years. In the Asia-Pacific region, China is a major contributor to the blueberry industry [2]. The rapid development of deep learning has resulted in many new types of agricultural equipment. As a large agricultural country, it is particularly important for China to improve modern agricultural technology so that agriculture can keep up to date with the pace of modernization, such as that in the blueberry industry. Therefore, it is extremely important to use deep learning technology to develop an automated blueberry picking system, which can not only reduce a lot of human and material resources consumed by traditional picking methods, but also reduce the waste of resources caused by untimely picking. As an integral part of the fruit and vegetable picking robot system, the visual recognition system plays a vital role in fruit and vegetable target recognition and positioning, automatic picking and fruit and vegetable yield estimation [3]. However, the accuracy of object detection is important for the location of blueberries with different maturity levels in the clustered blueberry [4]. Therefore, it is necessary to design a detection model that is suitable for specific crop picking. For the fruit detection model, the accuracy of the detection and the lightweight design of the model are the key aspects. This paper studies the problem from these two aspects, and the specific contributions are as follows:

1. A blueberry dataset is constructed. Blueberry images growing in the natural environment were collected, and three kinds of blueberry with different degrees of maturity were marked with the LabelImg software. The blueberry images were augmented by data augmentation technology to enhance the generalization of the model, which can effectively avoid the overfitting problem during the training process.
2. A lightweight blueberry recognition model based on multi-scale and attention fusion is proposed. Firstly, we design a new attention module, NCBAM, which is added to the backbone network for improving the feature extraction ability of the model. Secondly, the small target detection layer is added to improve the multi-scale recognition ability of blueberries. Finally, the C3Ghost module is introduced into the backbone network to facilitate the reduction in model parameters.
3. The proposed blueberry recognition model based on improved YOLOv5 is validated. Experiments show that it can effectively improve the recognition accuracy of blueberries, which is beneficial to the development of orchard automatic picking.

2. Related Work

Blueberries are widely planted because of rich nutrition and high value [3], but blueberries growing in the natural environment are usually dense and sticky; what is worse, they are prone to complex backgrounds such as shading of branches and leaves. Therefore, rapid and accurate identification of blueberries is currently very challenging. Using deep learning technology to design a blueberry recognition model with excellent performance is one of the key points to realize automatic picking system. Therefore, it is necessary to conduct in-depth research on it.

The detection speed of YOLOv5 is faster than that of YOLOv3 [5] and YOLOv4 [6], and it can more accurately detect targets in the case of complex backgrounds and occluded targets. Therefore, the current target detection is generally improved based on the YOLOv5 model [7–20]. In order to improve the detection performance of YOLOv5, the network is generally improved from three aspects: backbone network, neck network and prediction network. More details are as follows:

There are several ways to improve the backbone network. Yan et al. [7] replaced BottleneckCSP module in the backbone network of the original YOLOv5s with BottleneckCSP-2 module in order to effectively reduce the number of model parameters. Secondly, the SE (Squeeze-and-Excitation) module of the visual attention mechanism network is added to the backbone network to improve the expression ability of the model; Similarly, Chen et al. [8] also added an SE module to the backbone network for improving the sensitivity of the model to channel features. The proposed improved network model can effectively identify graspable apples that are not occluded or only occluded by leaves, and ungraspable apples that are occluded by branches or other fruits. In order to detect objects in the images with a complex background, Hu et al. [9] improved the C3 module in the backbone network using the convolution kernel group to enhance the feature extraction of the detected object and the attention module to focus on the whole object; Li et al. [10] replaced the ordinary convolution in the network model with the depthwise separable convolution, which reduced the number of network parameters and improved the detection accuracy of apple fruits. Luo et al. [11] proposed a new detection method named YOLOv5-Aircraft, which solved the problem of insufficient detection accuracy and slowed the detection speed of aircraft targets in remote sensing images under complex backgrounds. In the method, the hourglass-shaped module CSAndGlass is designed on the backbone feature extraction network of YOLOv5 and the original residual module is replaced by CSAndGlass, which reduces the semantic loss. Therefore, based on the above research, we can clearly know that in terms of backbone network improvement, firstly, adding an attention module can enhance the feature extraction ability of detected objects, thereby improving the overall detection performance of the model. Secondly, the size of the model can be reduced by using the lightweight module, so as to achieve the purpose of improving the speed. Therefore, in this study, the NCBAM attention module we designed was added to the backbone network

to improve the feature extraction ability of blueberries, and the C3 module was replaced with C3Ghost to reduce the model size.

In terms of improving the neck network, Zhao et al. [12] proposed an improved network structure by adding a micro-scale detection layer, setting an a priori anchor box, and adjusting the confidence loss function of the detection layer based on IoU. The improved YOLOv5 method can accurately detect wheat peaks in UAV images, solve the problem of ear error detection and omission detection caused by occlusion conditions, enhance the feature extraction ability of wheat ears, and improve the detection accuracy. Zhu et al. [13] designed a new feature fusion layer to capture shallow features of the small boulder and combined Convolutional Block Attention Module (CBAM) and Effective Channel Attention Network (ECA-Net) to integrate a new attention module, which is added to the neck network to highlight information helpful for boulder detection. Through the study of [12,13], we concluded that in the field of small object detection, adding a small detection scale can improve the detection accuracy of small objects. Therefore, a small-scale detection layer is added in this study to improve the detection accuracy of blueberries, because the blueberry target is small in some images.

3. Materials and Methods

3.1. Blueberry Image Collection

The purpose of this study is to identify blueberries grown in natural environment. Not only is the background of blueberry images in natural environment complex, but the blueberry object is also seriously disturbed by external factors such as branches and leaves [21]. Secondly, blueberries grow in clusters, and each cluster usually contains blueberries of different degrees of maturity. As for blueberries, they can be roughly divided into three categories according to the grade of maturity. The color of ripe blueberries, semi-ripe blueberries and immature blueberries corresponds to purple-red, light red and cyan, respectively [22]. Figure 1, the 6000 × 4000 pixels blueberry images were taken with a Canon EOS 80D handheld digital SLR camera in Xinjian District, Nanchang City, Jiangxi Province. This research serves for the realization of an automatic picking system. Therefore, in order to make the experimental data more accurate, when shooting blueberry images, we simulated the lens turning in the automatic picking mode, shooting blueberries from the front, side, top and bottom angles, respectively. A total of 1000 images were collected, and the collection types were front, side, overlap, occlusion, and adhesion, as shown in Figure 2.

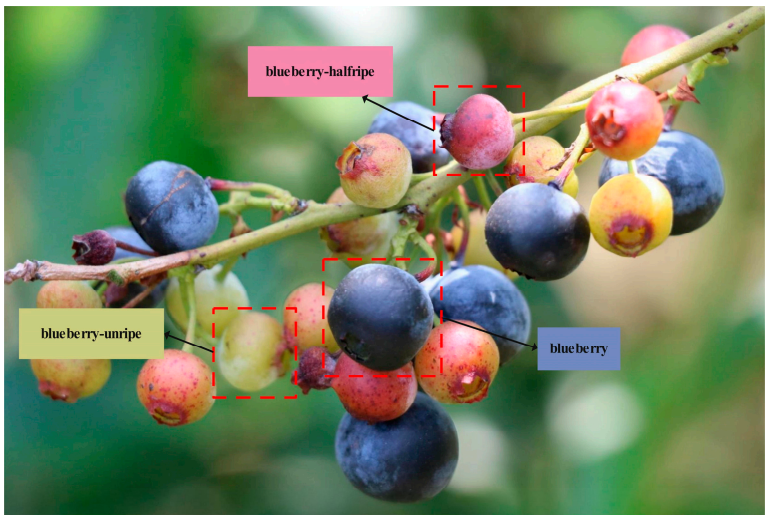
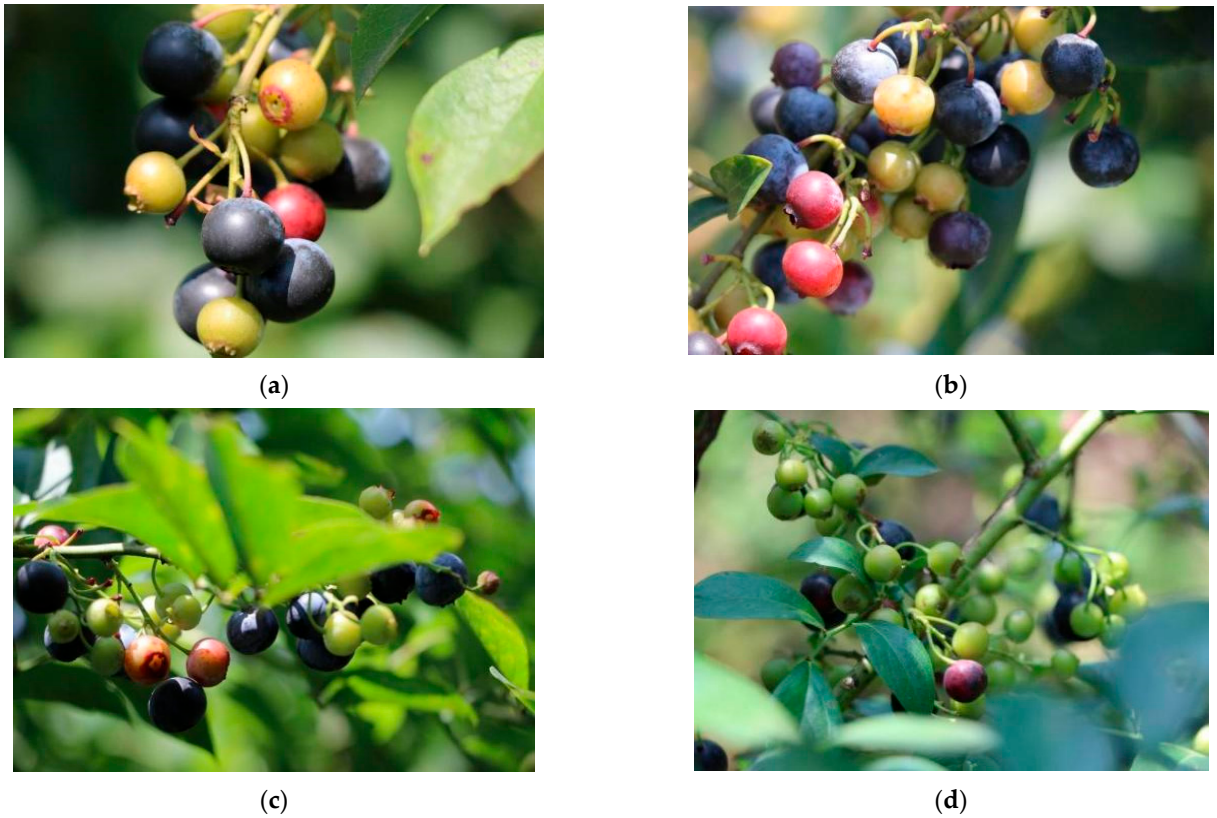


Figure 1. Blueberry ripeness comparison.



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

(**c**) (**d**)

Figure 2. Collection of pictures. (**a**) Dense adhesions; (**b**) severely occluded; (**c**) branches and leaves cover; (**d**) complex background.

3.2. Data Preprocessing

After collecting images, LabelImg software for manual annotation is used to complete the task of image labeling, which will generate an xml file containing the image name, size and position information of the blueberry for each image. We want to design a network model with high recognition accuracy for blueberry. However, the training of the neural network model often requires a large number of samples, but collecting a large number of samples needs a lot of human and material resources. Therefore, in order to obtain a neural network model with high recognition accuracy, it is necessary to use augmentation technology for expanding the dataset so as to enhance the generalization of the model, which can effectively avoid the problem of overfitting during the training process. Data enhancement methods include the following types: image enhancement, adding noise, adjusting brightness, rotation, translation, mirroring, and cropping. The dataset is expanded to 10,000 images by data augmentation. The number of labels in three categories is: 93,502 for ripe blueberries, 14,969 for semi-ripe blueberries, and 69,671 for unripe blueberries. The dataset is randomly divided according to the ratio of 6:2:2. After the division, there are 6000 training samples, 2000 verification samples and 2000 test samples, respectively (see Table 1). Remaining images are input into the network model as query images, and their annotation labels are used as the ground truth for loss calculation.

Table 1. Dataset partition.

Number of Pictures	Number of Train, Val, Test	Blueberry Ripeness	Number of Labels	Color
10,000 images	train set of 6000 images	blueberry blueberry-halfripe blueberry-unripe	55,502 8889 41,881	fuchsia light red green

Table 1. *Cont.*

Number of Pictures	Number of Train, Val, Test	Blueberry Ripeness	Number of Labels	Color
10,000 images	val set of 2000 images	blueberry blueberry-halfripe blueberry-unripe	18,120 2520 14,400	fuchsia light red green
	test set of 2000 images	blueberry blueberry-halfripe blueberry-unripe	19,880 3560 13,390	fuchsia light red green

4. The Proposed Method

4.1. YOLOv5 Original Network Structure

YOLOv5 is composed of input, backbone network, neck network and three detection heads. The overall structure of the original YOLOv5 network is shown in Figure 3. In this study, we use the newly released YOLOv5-6.1 version in 2022.

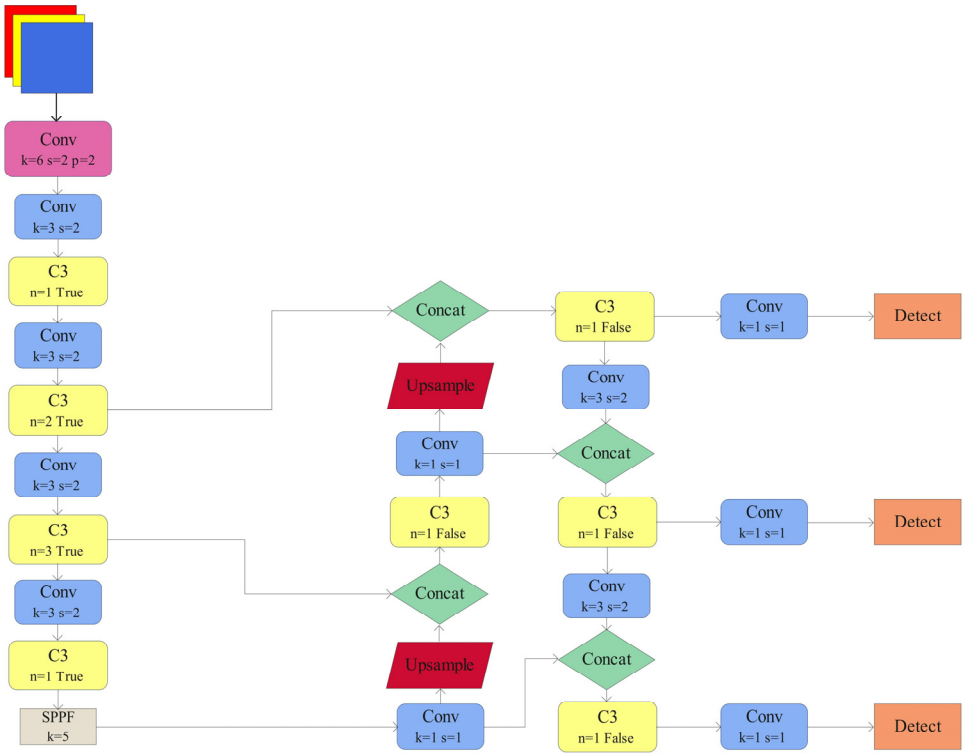


Figure 3. YOLOv5 network structure (The arrows indicate the direction of forward propagation).

(1) Input

Like Yolov4, Mosaic data augmentation is used in the YOLOv5 input [23]. In addition, YOLOv5 also has the function of adaptive anchor box calculation, which can adaptively calculate the best anchor box value according to different datasets.

(2) Backbone

Compared with the previous version, the backbone network part of YOLOv5 replaced the Focus module, which was the first layer of the network structure, with a convolution layer after the v6.0 version, and the convolution kernel size was 6 × 6, which is beneficial to reduce the amount of model parameters and improve the detection speed and accuracy. Figure 4 shows the Focus module used in the network before v6.0 version. The Focus module extracts pixels from high-resolution images and reconstructs them into low-resolution images, focusing on w and h dimension information and converting them into c-channel

dimensions, finally extracting different features through 3 × 3 convolution. In this way, the information loss caused by down-sampling can be reduced and receptive field can be increased. The 6 × 6 convolutional layer and the Focus module play the same role. However, currently in many GPU devices, the former is more efficient.

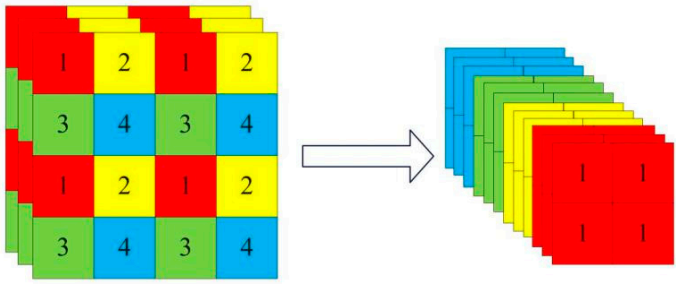


Figure 4. Structure diagram of focus (The arrow indicates slice operation).

Another improvement in the backbone network is to replace the original SPP module [24] with a faster SSPF module. The main function of SPPF is to extract and fuse high-level features by connecting three 5 × 5 MaxPooling operations in series, whose structure is shown in Figure 5.

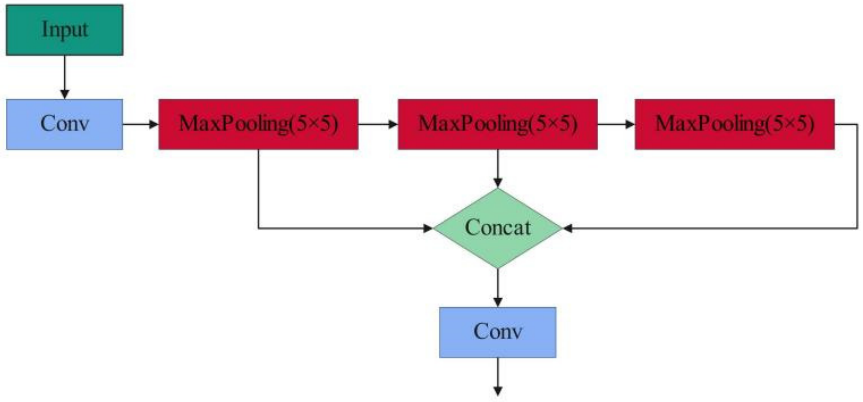


Figure 5. Structure diagram of SPPF (The arrows indicate the direction of forward propagation).

The role of C3 is to perform feature extraction on feature maps obtained from the previous convolutional layer.

(3) Neck

The neck network is adjacent to the backbone network, which fuses the top-level and bottom-level features by using the combination of FPN (Feature Pyramid Networks) + PAN (Path Aggregation Network) [25,26], as shown in Figure 6. The obtained feature maps of different sizes are input into the detection head to predict three objects of different sizes.

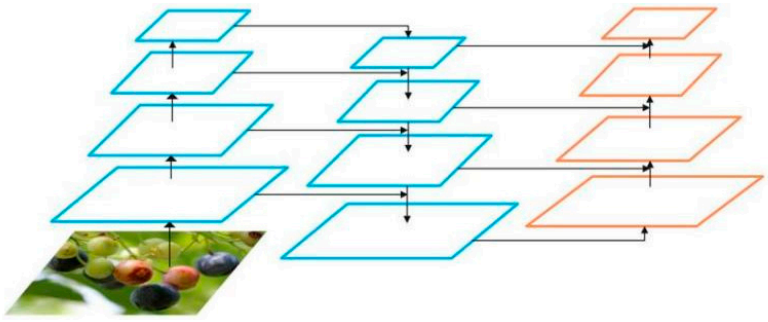


Figure 6. Structure diagram of FPN + PAN(The arrows indicate the direction of forward propagation).

(4) Prediction

The prediction network, which is consisted of three detection heads of different scales, predicts bounding box information and category information of the final output target based on the feature maps of different sizes output by the neck network.

(5) Loss function

CIOU_Loss is used as loss function [27]. CIOU_Loss exactly right adds a penalty term for the aspect ratio of the bounding box on the original basis, the consistency of the aspect ratio of the prediction box and the target box is very important.

4.2. Improved YOLOv5 Network Structure

Since our purpose is to obtain an excellent blueberry recognition model to identify blueberries with complex backgrounds, serious occlusion and dense adhesion, it is necessary to improve the original network structure of YOLOv5s to achieve the purpose of improving the detection accuracy so as to further realize the automatic picking of blueberries. The improved network is shown in Figure 7. Firstly, the NCBAM module we designed was added after each C3 module in the backbone network to improve the network model's ability, which is advantageous to extract blueberry features. Secondly, C3 module in the backbone network was replaced with C3Ghost to improve the detection speed of the model and reduce the weight of the model. Finally, in order to improve the ability of detecting individual small blueberries, the original three detection scales were expanded to four for realizing multi-scale blueberry recognition. Experiments show that our model can effectively identify blueberries, and the mAP is 2.4% higher than that of the original network.

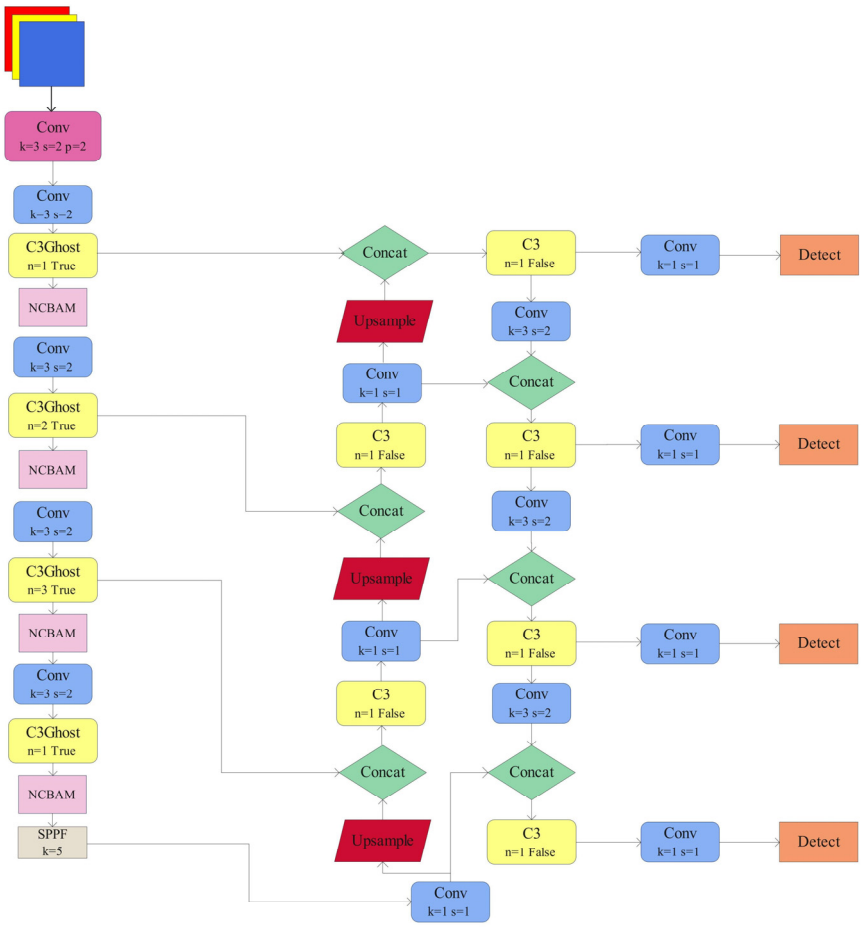


Figure 7. Improved YOLOv5 network structure (The arrows indicate the direction of forward propagation).

4.2.1. New Convolutional Block Attention Module

In this paper, we design a new attention module named NCBAM (New Convolutional Block Attention Module), as shown in Figure 8. The NCBAM, which focuses on features from channel and space, is composed of two branches, one branch connecting channel attention module and spatial attention module in parallel, and the other connecting channel attention module and spatial attention module in series [28]. In the improvement of the backbone network, an NCBAM is added after each C3 module in the backbone network to focus on the important features of blueberries and suppress unnecessary features, thereby improving the network recognition accuracy.

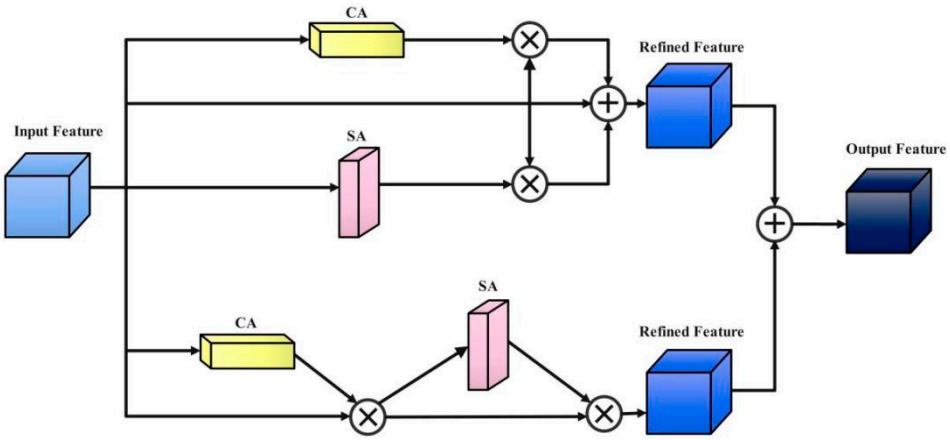


Figure 8. New Convolutional Block Attention Module (CA is the Channel Attention Module, SA is the Spatial Attention Module. The arrows indicate the direction of forward propagation. ⊗ represent the multiplication of feature map).

4.2.2. Small Object Detection Layer

There are three scale layers in the original YOLOv5 network, which detect feature maps of different scales obtained from the neck network, and down-samples the input image dimensions by a factor of 32, 16 and 8, respectively. However, the small-scale detection layer of YOLOv5 is less suitable for blueberries, because blueberries are not only small but also densely distributed in many images. Therefore, a new micro-scale detection layer is constructed to expand three scales of the original YOLOv5 to four. The new detection layer down-samples the input image dimension by 4 times and generates feature maps through fusing low-level spatial features and deep semantic features. Experiments show that the detection accuracy is improved after adding the detection layer, so the added micro-scale detection layer is suitable for detecting densely growing blueberries.

4.2.3. C3Ghost

At present, a large number of convolutional neural networks rely on stacked convolution to obtain feature maps for the purpose of improving network accuracy, which leads to huge network parameters and a large amount of calculation. Naturally, many modules are designed to reduce the computational load and improve the accuracy of the network, such as the Ghost module [29] used in our research, which is also a lightweight network module. Its design principle is that there is often redundant information in the feature map, which is of great significance to the overall consistency of input data and may become an important part of model optimization. GhostNet does not delete redundant information but uses a low-cost calculation to obtain these redundant feature map information, which improves the calculation accuracy and reduces the parameters to realize lightweight design. The model volume has increased after adding the NCBAM module and the small-scale detection layer, so we replaced all the C3 modules of the backbone network with the C3Ghost module to reduce the model volume while remaining the blueberry recognition accuracy. The C3Ghost structure is shown in Figure 9.

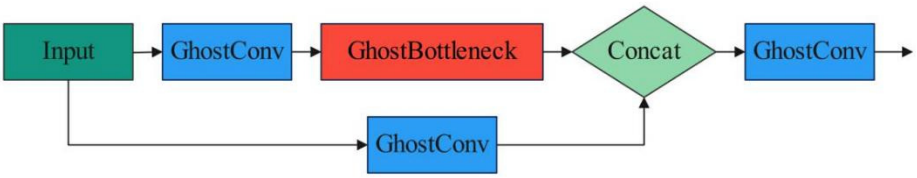


Figure 9. Structure diagram of C3Ghost (The arrows indicate the direction of forward propagation).

5. Results and Analysis

5.1. Lab Environment

The development framework of these experiments is PyTorch 1.7, the programming language is Python, and the computer configuration is: NVIDIA RTX 2080 Ti graphics card, Intel i7-9700k CPU @3.60 GHz, video memory 11 GB, and 64 GB memory Windows 10 system. The experimental environment configuration is shown in Table 2.

Table 2. Experimental environment.

Name	Related Configuration
Graphics Card	NVIDIA GeForce RTX 2080 Ti
Processor	Intel Core i7-9700K CPU @ 3.60 GHz
Memory	64 GB
System	Windows 10
Development Framework	PyTorch
Programming Language	Python

5.2. Model Evaluation Metrics

The evaluation of model performance in the field of target detection usually uses three types of evaluation indicators: Precision, Recall, and mAP (mean Average Precision) which is a combination of the first two. Precision represents the total amount of retrieved information, that is, the proportion of positive samples in all samples in the detection results. The calculation formula is Formula (1). Recall represents the proportion of the number of positive samples in the number of labeled positive samples, and the calculation formula is Formula (2). The mAP is the average accuracy rate. In the detection of multiple categories of objects, each category can draw a curve according to P and R, AP is the area under the curve, and mAP is the average of multiple categories. The calculation formula is Formula (3).

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{1}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} AP = \int_0^1 P(R)dR \tag{3}$$

TP (True Positive) indicates that both the detection result and the true value are blueberries; FP (False Positive) indicates the number of samples marked as false but detected as positive samples, that is, the number of false blueberries detected; FN (False Negative) indicates that the number of samples marked as positive but predicted as negative classes, that is, the number of missed detection blueberries. In this experiment, the larger the mAP value, the better the algorithm detection effect and the higher the model recognition performance.

5.3. Experimental Results and Analysis

5.3.1. Determination of Basic Network Structure

At present, YOLOv5 has released five versions according to different network widths and depths. In order to find the most suitable model for this study, we conduct experiments

with network structures of different widths and depths on the self-built blueberry dataset. The training results of YOLO-v5 with different depths and widths are shown in Table 3. According to the experimental results, the mAP of YOLO-v5n is 80.8%, the mAP of YOLO-v5x is 82.8%, Although the network model with increased width and depth has a small improvement in the detection performance of blueberries, the weight and volume of the model are also relatively large. The model parameters of YOLO-v5x is 86186872, which is 98% larger than that of YOLO-v5n. YOLO-v5n is the simplest network, but the recall of YOLO-v5n is 72.9%, which is an obvious drop compared to other network models. The mAP of the YOLO-v5s network is 80.8%, which is little difference in accuracy with other models, and the model parameters is 7,018,216, which is 92% less than the largest YOLO-v5x model. Therefore, considering comprehensively, this study uses YOLO-v5s as the basic network structure.

Table 3. Performance comparison of network structures of different widths and depths.

YOLO-v5	Precision (%)	Recall (%)	mAP (%)	Parameters
YOLO-v5n	83.9	72.9	80.8	1,763,224
YOLO-v5s	83.1	75.8	80.8	7,018,216
YOLO-v5m	83.3	75.2	81.6	20,861,016
YOLO-v5l	82.5	76.1	81.7	46,119,048
YOLO-v5x	84.6	76.7	82.8	86,186,872

5.3.2. Validation of the Effectiveness of Proposed Attention Module

In order to verify which attention mechanism has the best effect in this study, some comparative experiments were conducted between the proposed attention module and two popular attention modules, Coordinate Attention (CA) and the Squeeze-and-Excitation (SE) attention mechanism, respectively. We place three different attention mechanisms in the same position of the YOLOv5s network architecture for comparison, and the experimental results are shown in Table 4. The experimental results show that our proposed NCBAM has the greatest effect on this study, and the mAP is improved by 0.9% compared to the original YOLOv5s. Therefore, this study selects NCBAM as the attention module of the backbone network.

Table 4. Comparison results of different attention modules.

Attention Module	Precision (%)	Recall (%)	mAP (%)
CA	83.9	75.6	81.3
SE	84.4	75	80.8
NCBAM	83.8	76.1	81.7

5.3.3. Model Lightweight Validity Verification

In order to verify the effectiveness of C3Ghost in reducing the weight of the model, we performed the following experiments based on our improved NCBAM attention mechanism and four-layer network. The experimental results are shown in Table 5. The parameters of our improved network model are decreased by 15%, and mAP is increased by 0.1% more than the model without C3Ghost module.

Table 5. Comparison of parameters and performance after C3Ghost module replacement.

	Parameters	mAP (%)
YOLOv5s + NCBAM + Four	7722120	83.1
YOLOv5s + NCBAM + Four + C3Ghost	6559632	83.2

5.3.4. Validation of the Improved Method

In order to verify the effectiveness of the improvement of YOLOv5 in this study, we performed a comparative experiment of different improvement methods. The experimental

results are shown in Table 6. The mAP value is increased by 0.9% after adding the NCBAM attention mechanism, which proves that NCBAM has a certain effect on the feature extraction of blueberries. After adding a small-scale detection layer, mAP reaches 83.1%, which improves the ability to detect blueberries at multi-scale. After adding C3Ghost, the model reduces the number of parameters while ensuring the accuracy. Overall, our proposed method improves the mAP value of the original YOLOv5s by 2.4%.

Table 6. Experimental results after different modules being added (① represents NCBAM, ② represents Small object detection layer, and ③ represents C3Ghost).

	Blueberry AP (%)	**Blueberry-Halfripe AP (%) AP**	**Blueberry-Unripe AP (%)**	**mAP (%)**
YOLOv5s	90.5	72.3	79.5	80.8
①	90.3	75.3	79.6	81.7
① + ②	92.1	75.1	82.2	83.1
① + ② + ③	91.9	75	82.8	83.2

5.3.5. Comparison with Other Methods

To evaluate the performance of our improved YOLOv5 model, we compare it with the following common convolutional neural networks, RetinaNet [30], Yolov3, MobileNetv3-YOLOv5 [31], YOLOv5. The experimental results are shown in Table 7, from which it is concluded that our model outperforms several other models. Additionally, it is 2.4% higher than the mAP of the original YOLOv5, which proves the effectiveness of our model.

Table 7. Performance comparison on different network models.

Network Model	**mAP (%)**
RetinaNet	71.5
Yolov3	80
MobileNetv3-YOLOv5	79.6
YOLOv5s	80.8
Ours	83.2

6. Conclusions

This study improves the YOLOv5 network and adds our newly designed NCBAM to the backbone network for improving the model's ability to extract blueberry features. Then, the C3 module in the backbone network is replaced with the C3Ghost module in order to reduce the model parameters. Finally, a small target detection layer is added to detect blueberry at multiple scales, and the ability to identify blueberries is improved. It can be seen from the experimental results that the improved network has a 2.4% increase in mAP compared with the original YOLOv5 network, which proves that the improved model can effectively improve the recognition accuracy of blueberries. In addition, it can also be used detect three kinds of blueberry of different maturity, providing an accurate blueberry positioning for automatic blueberry picking system, thereby reducing economic losses caused by untimely manual picking, improving the economic benefits of the blueberry industry, and promoting the development of fruit and vegetable picking robot systems. Compared with the original YOLOv5 network, this model has more network parameters. In the next work, we will continue to research on reducing the network parameters and improving the detection ability.

Author Contributions: X.M. conceived the paper, designed and conducted experiments, and wrote the paper. W.Y. provides guidance for thesis innovation and guides thesis revision. W.H., software. P.T. provided constructive comments on the research and revised the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 61462038; the Natural Science Foundation of Jiangxi Province, grant number 20212BAB212005; Open Project of State Key Laboratory of Zhejiang University, grant number A2029; the Natural Science Foundation of Jiangxi Province, grant number 20212BAB202020.

Data Availability Statement: The blueberry dataset is at https://github.com/mxx0118/Blueberry-Datasets (accessed on 23 September 2022).

Acknowledgments: The authors would like to thank the anonymous reviewers for their critical comments and suggestions for improving the manuscript.

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

References

1. Li, Y.D.; Pei, J.B.; Sun, H.Y. Status and Prospect of Global Blueberry Industry. *J. Jilin Agric. Univ.* **2018**, *40*, 421–432.
2. Li, Y.D.; Sun, H.Y.; Chen, L. Report on the development of China's blueberry industry. *China Fruit Tree* **2016**, *5*, 1–10.
3. Zhang, F.; Chen, Z.J.; Bao, R.F.; Zhang, Z.C.; Wang, Z.H. Recognition of dense cherry tomatoes based on improved YOLOv4-LITE lightweight neural network. *Trans. Chin. Soc. Agric. Eng. Trans. CSAE* **2021**, *37*, 270–278.
4. Lu, F.; Liu, H.H.; Huang, C.Y.; Yang, Y.; Xie, Y.; Liu, C.X. Overview on Deep Learning-Based Object Detection. *Comput. Syst. Appl.* **2021**, *30*, 1–13.
5. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
6. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
7. Yan, B.; Fan, P.; Lei, X.Y.; Liu, Z.J.; Yang, F.Z. A Real-Time Apple Targets Detection Method for Picking Robot Based on Improved YOLOv5. *Remote Sens.* **2021**, *13*, 1619. [CrossRef]
8. Chen, Z.Y.; Wu, R.H.; Lin, Y.Y.; Li, C.Y.; Chen, S.Y.; Yuan, Z.N.; Chen, S.W.; Zou, X.J. Plant Disease Recognition Model Based on Improved YOLOv5. *Agronomy* **2022**, *12*, 365. [CrossRef]
9. Hu, G.S.; Wu, J.T.; Bao, W.X.; Zeng, W.H. Detection of Ectropis oblique in complex background images using improved YOLOv5. *Trans. Chin. Soc. Agric. Eng. Trans. CSAE* **2021**, *37*, 191–198.
10. Li, Z.J.; Yang, S.H.; Shi, D.S.; Liu, X.X.; Zheng, Y.J. Yield estimation method of apple tree based on improved lightweight YOLOv5. *Smart Agric.* **2021**, *3*, 100–114.
11. Luo, S.; Yu, J.; Xi, Y.J.; Li, X. Aircraft Target Detection in Remote Sensing Images Based on Improved YOLOv5. *IEEE Access* **2022**, *10*, 5184–5192. [CrossRef]
12. Zhao, J.Q.; Zhang, X.H.; Yan, J.W.; Qiu, X.L.; Yao, X.; Tian, Y.C.; Zhu, Y.; Cao, W.X. A Wheat Spike Detection Method in UAV Images Based on Improved YOLOv5. *Remote Sens.* **2021**, *13*, 3095. [CrossRef]
13. Zhu, L.L.; Geng, X.; Li, Z.; Liu, C. Improving YOLOv5 with Attention Mechanism for Detecting Boulders from Planetary Images. *Remote Sens.* **2021**, *13*, 3776. [CrossRef]
14. Wang, F.H.; Sun, Z.X.; Chen, Y.; Zheng, H.; Jiang, J. Xiaomila Green Pepper Target Detection Method under Complex Environment Based on Improved YOLOv5s. *Agronomy* **2022**, *12*, 1477. [CrossRef]
15. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5:Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Online, 11–17 October 2021; pp. 2778–2788.
16. Zhao, Z.Y.; Yang, X.X.; Zhou, Y.C.; Ge, Z.D.; Liu, D.F. Real-time detection of particleboard surface defects based on improved YOLOV5 target detection. *Sci. Rep.* **2021**, *11*, 21777. [CrossRef]
17. Yao, J.; Qi, J.M.; Zhang, J.; Shao, H.M.; Yang, J.; Li, X. A real-time detection algorithm for Kiwifruit defects based on YOLOv5. *Electronics* **2021**, *10*, 1711. [CrossRef]
18. Guan, Z.X.; Liu, H.; Zuo, Z.J.; Pan, L.B. Design a Robot System for Tomato Picking Based on YOLO v5. *IFAC-PapersOnLine* **2022**, *55*, 166–171.
19. Li, R.; Wu, Y.P. Improved YOLO v5 Wheat Ear Detection Algorithm Based on Attention Mechanism. *Electronics* **2022**, *11*, 1673. [CrossRef]
20. Zhang, P.; Liu, X.M.; Yuan, J.; Liu, C.L. YOLO5-spear: A robust and real-time spear tips locator by improving image augmentation and lightweight network for selective harvesting robot of white asparagus. *Biosyst. Eng.* **2022**, *218*, 43–61. [CrossRef]
21. MacEachern, C.B.; Esau, T.J.; Schumann, A.W.; Hennessy, P.J.; Zaman, Q.U. Detection of Fruit Maturity Stage and Yield Estimation in Wild Blueberry Using Deep Learning Convolutional Neural Networks. *Smart Agric. Technol.* **2022**, *3*, 100099. [CrossRef]
22. Wang, L.S.; Qin, M.X.; Lei, J.Y.; Wang, X.F.; Tan, K.Z. Blueberry maturity recognition method based on improved YOLOv4-Tiny. *Trans. Chin. Soc. Agric. Eng. Trans. CSAE* **2021**, *37*, 170–178.
23. Wang, H.; Song, Z.L. Improved Mosaic: Algorithms for more Complex Images. *J. Phys. Conf. Ser.* **2020**, *1684*, 012094.
24. Huang, Z.C.; Wang, J.L.; Fu, X.S.; Yu, T.; Guo, Y.Q.; Wang, R.T. DC-SPP-YOLO: Dense connection and spatial pyramid pooling based YOLO for object detection. *Inf. Sci.* **2020**, *522*, 241–258. [CrossRef]

25. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, United States Hawaii Convention Center, Honolulu, HI, USA, 19 April 2017; pp. 2117–2125.
26. Wang, W.H.; Xie, E.Z.; Song, X.G.; Zang, Y.H.; Wang, W.J.; Lu, T.; Yu, G.; Shen, C.H. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea (South), 16 August 2019; pp. 8440–8449.
27. Zheng, Z.H.; Wang, P.; Liu, W.; Li, J.Z.; Ye, R.G.; Ren, D.W. Distance-IoU loss: Faster and better learning for bounding box regression. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 12993–13000. [CrossRef]
28. Woo, S.; Park, J.; Lee, J.Y.; Kweon, S. Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
29. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1580–1589.
30. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
31. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.X.; Wang, W.J.; Zhu, Y.K.; Pang, R.M.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF international conference on computer vision, Seoul, Korea (South), 16 August 2019; pp. 1314–1324.