agriculture

Article

Optimizing the YOLOv7-Tiny Model with Multiple Strategies for Citrus Fruit Yield Estimation in Complex Scenarios

Juanli Jing [1], Menglin Zhai [1], Shiqing Dou [1,*], Lin Wang [1], Binghai Lou [2], Jichi Yan [3] and Shixin Yuan [1]

[1] College of Geomatics and Geoinformation, Guilin University of Technology, Guilin 541006, China; 2003080@glut.edu.cn (J.J.); 2120222042@glut.edu.cn (M.Z.); 2120211899@glut.edu.cn (L.W.); 1020232072@glut.edu.cn (S.Y.)
[2] Guangxi Academy of Specialty Crops, Guilin 541004, China; binghai.lou@hotmail.com
[3] College of Mechanical and Control Engineering, Guilin University of Technology, Guilin 541006, China; 2019178@glut.edu.cn
* Correspondence: doushiqing@glut.edu.cn; Tel.: +86-18278367609

Abstract: The accurate identification of citrus fruits is important for fruit yield estimation in complex citrus orchards. In this study, the YOLOv7-tiny-BVP network is constructed based on the YOLOv7-tiny network, with citrus fruits as the research object. This network introduces a BiFormer bilevel routing attention mechanism, which replaces regular convolution with GSConv, adds the VoVGSCSP module to the neck network, and replaces the simplified efficient layer aggregation network (ELAN) with partial convolution (PConv) in the backbone network. The improved model significantly reduces the number of model parameters and the model inference time, while maintaining the network's high recognition rate for citrus fruits. The results showed that the fruit recognition accuracy of the modified model was 97.9% on the test dataset. Compared with the YOLOv7-tiny, the number of parameters and the size of the improved network were reduced by 38.47% and 4.6 MB, respectively. Moreover, the recognition accuracy, frames per second (FPS), and F1 score improved by 0.9, 2.02, and 1%, respectively. The network model proposed in this paper has an accuracy of 97.9% even after the parameters are reduced by 38.47%, and the model size is only 7.7 MB, which provides a new idea for the development of a lightweight target detection model.

Keywords: citrus identification; computer vision; YOLOv7-tiny; deep learning; load estimation

check for updates

Citation: Jing, J.; Zhai, M.; Dou, S.; Wang, L.; Lou, B.; Yan, J.; Yuan, S. Optimizing the YOLOv7-Tiny Model with Multiple Strategies for Citrus Fruit Yield Estimation in Complex Scenarios. *Agriculture* **2024**, *14*, 303. https://doi.org/10.3390/agriculture14020303

Academic Editors: Xiuguo Zou, Xiaochen Zhu, Wentian Zhang, Yan Qian and Yuhua Li

Received: 1 January 2024
Revised: 7 February 2024
Accepted: 9 February 2024
Published: 13 February 2024

1. Introduction

In the last decade, China has been a world leader in citrus fruit production [1], which has become an important factor in the economic growth of agricultural cultivation. In citrus fruit production, preharvest yield estimation is an essential tool for assessing fruit quality and planting techniques and for reflecting the market supply and demand; moreover, this method is crucial for guiding agricultural production [2]. Traditional methods for estimating production usually involve manual sampling, which is inefficient and labor intensive. Currently, deep learning technology has shown excellent advantages in image detection, and the fast and accurate detection ability of a target detection network can provide a more efficient means for citrus fruit yield estimation and information counting.

In recent years, object detection has played an important role in the field of agriculture. Sozzi et al. [3] used a detection network to determine grape yield estimation. Wang et al. [4] employed the YOLOv5 network to detect small apples for early yield estimation. Cardellicchio et al. [5] used the YOLOv5 detection network to study the phenotypic characteristics of tomato plants, so that tomato fruits and flowers could be accurately identified. In addition, the target detection networks have also demonstrated a good detection performance in fruit identification and counting [6–10]. Bi et al. [11] used migration learning to train the FastR-CNN model for citrus fruit recognition with an average accuracy of 86.6%. Chen et al. [12]

improved the YOLOv7 network for citrus fruit detection by introducing a small-target detection layer and lightweight convolution with a model size of 24.26 M. However, the above model suffers from an excessive number of parameters and computational effort. Therefore, the lightweight YOLO series of inspection networks is the preferred choice [13–15]. Lightweight networks can increase the inference speed and reduce the model size through cutting the depth and width of the model. Huang et al. [16] improved upon the YOLOv5 lightweight model via introducing the CBAM attention mechanism, which has an average recognition accuracy of 91.3% for citrus fruits. Wang et al. [17] used the YOLOv5 model by introducing migration learning to detect apples with a model size of 4.07 M, but the detection accuracy was only 83.1%. Ma et al. [18] introduced a BiFPN module based on YOLOv7-tiny for detecting apples under different weather conditions, with an accuracy of 80.1%. Wang et al. [19] replaced the regular convolution in the YOLOv7-tiny backbone network by using variability convolution and introducing an SE attention mechanism for detecting millet chili at different maturity levels, with a model accuracy of 90.3%. In the above studies, most of the networks focused on local area feature extraction and matching, and had a lightweight character while losing the semantic information of the target fruits. Thus, they are prone to both missing and misdetecting target fruits when performing a multiscale target detection task, resulting in a lower accuracy in complex scenarios or small-target contexts. Therefore, to further improve the network's ability to detect citrus plants in complex environments and to meet the current demand for lightweight edge devices, there is an urgent need to explore in-depth strategies and methods to further enhance the detection capability of lightweight YOLO networks.

The purpose of this study is to solve the problems of redundant parameters and the high computation of target detection networks in complex scenarios. We propose a lightweight and high-precision network, YOLOv7-tiny-BVP, through module replacement and parameter optimization of reconfigurable network layers. The network includes the BiFormer attention mechanism, which improves the capture of spatial and contextual semantic information about citrus plants via capturing bidirectional dependencies in the input sequence. Moreover, regular convolution is replaced with GSConv in the neck network, and the VoVGSCSP module is introduced to improve detection accuracy and to reduce the model inference time. In the backbone network, the simplified ELAN is replaced with PConv, which leads to a more efficient feature extraction and reduces the number of network computations. The network is lightweight while maintaining high performance, which opens up the possibility of future applications of lightweight networks.

2. Materials and Methods

2.1. Data Acquisition

The citrus fruit images were collected at the Pengyu Brothers Citrus Demonstration Base in Gongcheng County, Guilin City, Guangxi Zhuang Autonomous Region. The data were collected on 11 November 2022, at 2:00–5:00 p.m. In this study, to improve sample representativeness, model applicability, and device compatibility, images were acquired from the Huawei Mate40 Pro phone and Nova 7 phone. To ensure the representativeness of the sample data, we obtain citrus fruit images at different distances varying from 0.2 to 2.0 m away from the fruit trees, and various real growth conditions were taken into account, such as upward view, downward view, downlight, backlight, fruit sparseness, denseness, shading, overlapping, etc. Finally, a total of 525 high-definition citrus fruit images with a resolution of 4096 × 3072 were captured in JPG format.

2.2. Data Enhancement

In the natural environment, there are a variety of disturbing factors, such as plant stems, leaves, and light. To improve the robustness of the model and prevent any overfitting caused by too little training data [20], 182 representative images were selected for data augmentation in this study. First, the original images were manually labeled using the image annotation software LabelImg. Then, the mixup method, the mosaic method, image equalization, gray scaling, and gamma transformation were used for mixed data enhancement. After manual screening, 3182 data points were ultimately obtained. The effect after data enhancement is shown in Figure 1.

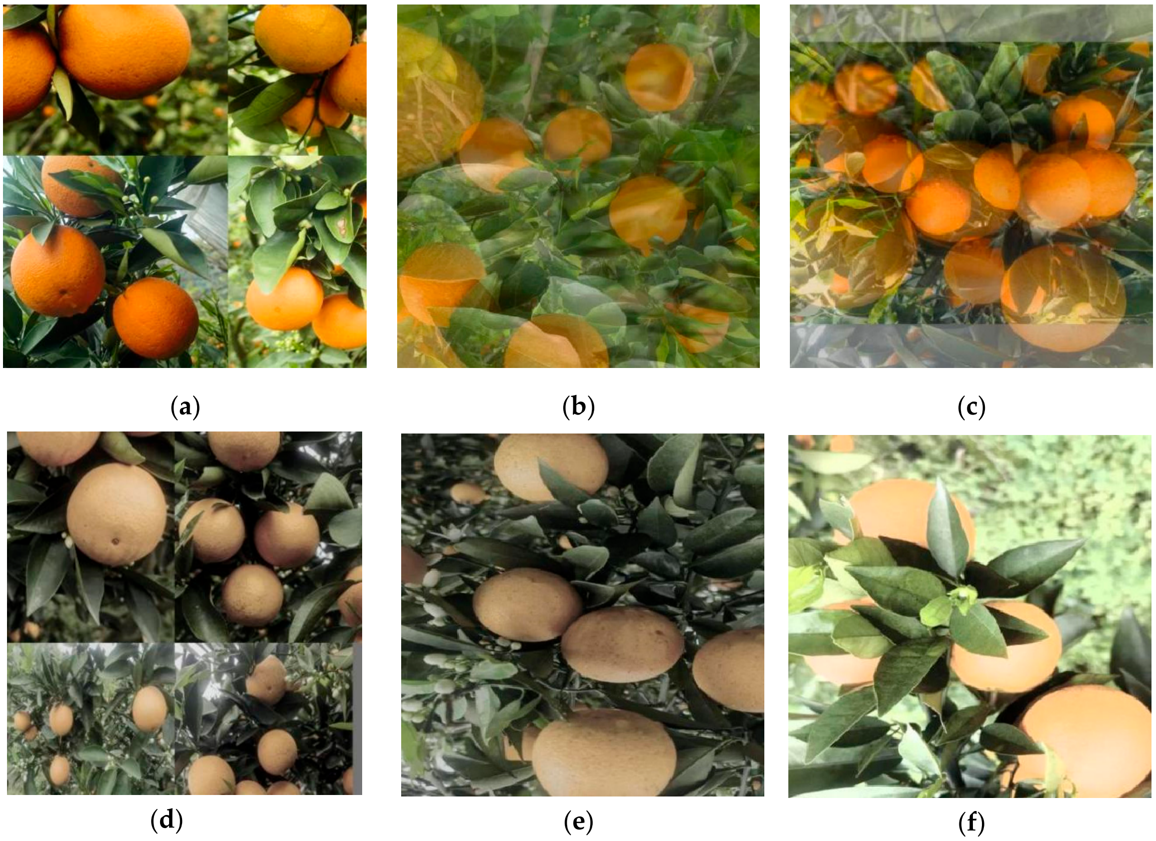


Figure 1. Data enhancement chart. (**a**) Mosaic enhancement; (**b**,**c**) Mixup enhancement; (**d**) Mosaic enhancement and gray scaling; (**e**) Gray scaling; (**f**) Gamma transform.

2.3. Dataset Preparation

In this study, the data were stored according to the format of Microsoft's publicly available dataset MS COCO, and the dataset was randomly divided into training, validation, and testing sets at a ratio of 7:2:1. The dataset contained 3182 images, including 182 original citrus fruit images and 3000 enhanced images, and 39,483 citrus fruits were labeled. The training set consisted of 2290 images that labeled 28,279 citrus fruits. The validation set consisted of 573 images in which 7181 fruits were labeled. The remaining 319 images comprised the test set, which included 4023 citrus fruits. According to the COCO dataset standard, citrus fruit targets with a resolution of less than 32 pixels × 32 pixels in the image were defined as small targets in this experiment. Table 1 demonstrates the division of the dataset.

Table 1. Dataset segmentation.

Dataset	Proportions	Images	Number of Captures
Training dataset	70%	2290	28,279
Validation dataset	20%	573	7181
Test dataset	10%	319	4023
Full data	100%	3182	39,483

2.4. Test Platforms and Parameters

This experiment was based on Windows 10 with the following hardware configuration: an i7-9800X@3.80 GHz CPU, 32 GB of RAM, and an Nvidia GeForce RTX2060 graphics card. We used PyCharm as the IDE, Python version 3.9 as the compiler, and PyTorch 1.17 as the test framework. In addition, CUDA version 12.1 parallel computing was used in combination with the cuDNN version 11.7 deep neural network acceleration library, and OpenCV 4.3.5 was selected as the image processing library.

To ensure the effectiveness of the model, we set the parameters of the initialized model as follows: begin training from 0 without loading the official pretraining weights, fix the random seeds, and accelerate the model training process by means of adjusting the number of threads used by CPU during data loading. The network hyperparameters were configured as follows: the input image size was 640 pixels × 640 pixels, and the batch size was 32. For model optimization, the stochastic gradient descent (SGD) method was used by setting the initial learning rate, momentum factor, and weight decay factor to 0.001, 0.937, and 0.0005, respectively. The model was trained for a total of 300 epochs, the weights were saved every 10 epochs, and the weights with the highest accuracy were ultimately selected for validation and testing.

In this experiment, COCO dataset-related metrics were used to evaluate the model performance. The evaluation metrics included precision (P), recall (R), mean average precision (mAP), F1 score (F1), model parameters, model size, and FPS, and the coefficient of determination R^2 was selected as the evaluation indicator of the prediction model. The equations for the aforementioned assessment indicators are shown in (1)–(4):

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

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

$$mAP = \frac{\Sigma_1^N \int_0^1 P(R)dR}{N} \tag{3}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{4}$$

In the above equations, TP denotes the positive sample predicted by the model, referring to the number of citrus fruits correctly identified, FP denotes the negative sample predicted by the model, referring to the number of citrus fruits incorrectly identified, and FN denotes citrus fruits that are not correctly detected.

3. Multi-Strategy Construction of a New YOLOv7-Tiny-BVP Network

3.1. YOLOv7-Tiny Network Infrastructure

The YOLOv7 target detection network is a detector proposed by Alexey Bochkovskiy's team in July 2022. On the Microsoft COCO public dataset, YOLOv7 outperforms all currently known detectors in terms of both speed and accuracy. Compared to its predecessors, the authors innovatively proposed the composite model scaling method and the extended efficient layer aggregation network (E-ELAN) to increase the depth and width of the network, which improved the feature expressiveness and detection performance of the network. Moreover, methods such as the dynamic label assignment strategy are used to improve the inference speed and detection accuracy [21].

YOLOv7-tiny is a lightweight network of the YOLOv7 series that consists of three main parts: the backbone, the neck, and the head; its network structure is shown in Figure 2. To achieve the lightweight standard, YOLOv7-tiny reduces the number of convolutions in the ELAN module, MP module, and SPPCSPC module and uses regular convolution instead of the CBL module, but does not fully integrate the advantageous modules in the YOLOv7 structure, which weakens the feature learning ability of the network to a certain extent [22].

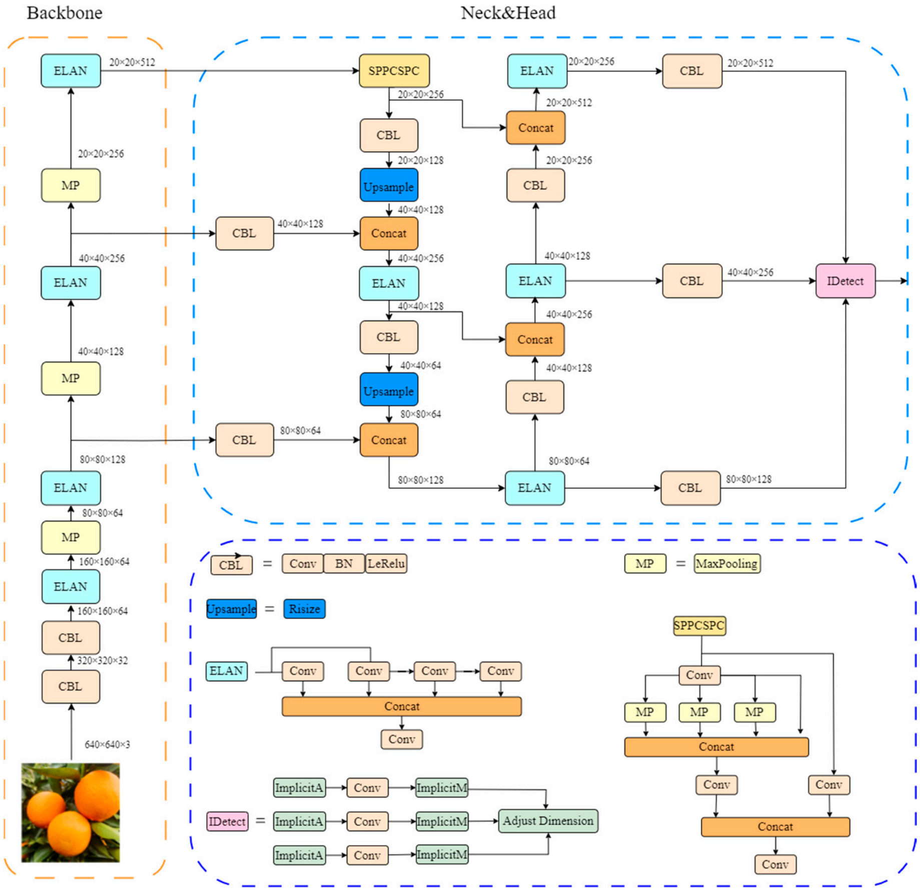


Figure 2. Diagram of the YOLOv7-tiny network.

3.2. Construction of a New YOLOv7-Tiny-BVP Network Using Multi-Strategy

This study explored methods for improving the new network based on the YOLOv7-tiny model. That is, the BiFormer attention mechanism is used after the SPPCSPC module, which enhances the network's fruit feature extraction ability in complex scenes; second, in the neck and head part, the regular convolution is replaced by GSConv, and the VoVGSCSP module is introduced, improving the detection accuracy and reduces the model inference time, simultaneously. Finally, the PConv module is used to replace the simplified ELAN structure in the backbone, which can reduce the network parameters and redundant computations, while maintaining detection accuracy. The new network named YOLOv7-tiny-BVP is constructed through multipolicy fusion; the structure is shown in Figure 3.

3.2.1. BiFormer Dynamic Sparse Attention Mechanism for Dual-Layer Routing

The BiFormer is an improvement and extension of the transformer model and consists of an encoder and a decoder. The encoder is used to encode the input sequence, and the decoder is used to generate the output sequence [23]. The BiFormer structure, which introduces a new bidirectional attention mechanism, is shown in Figure 4. The attention scores of each position are calculated via inputting the query vector (Q), key vector (K), and value vector (V) in the sequence, thereby establishing a global association within the sequence [24]. In citrus fruit recognition and detection tasks, contextual information is crucial for locating targets. The BiFormer can more comprehensively capture the background information and contextual relationships around citrus fruit plants, improving fruit detection accuracy.

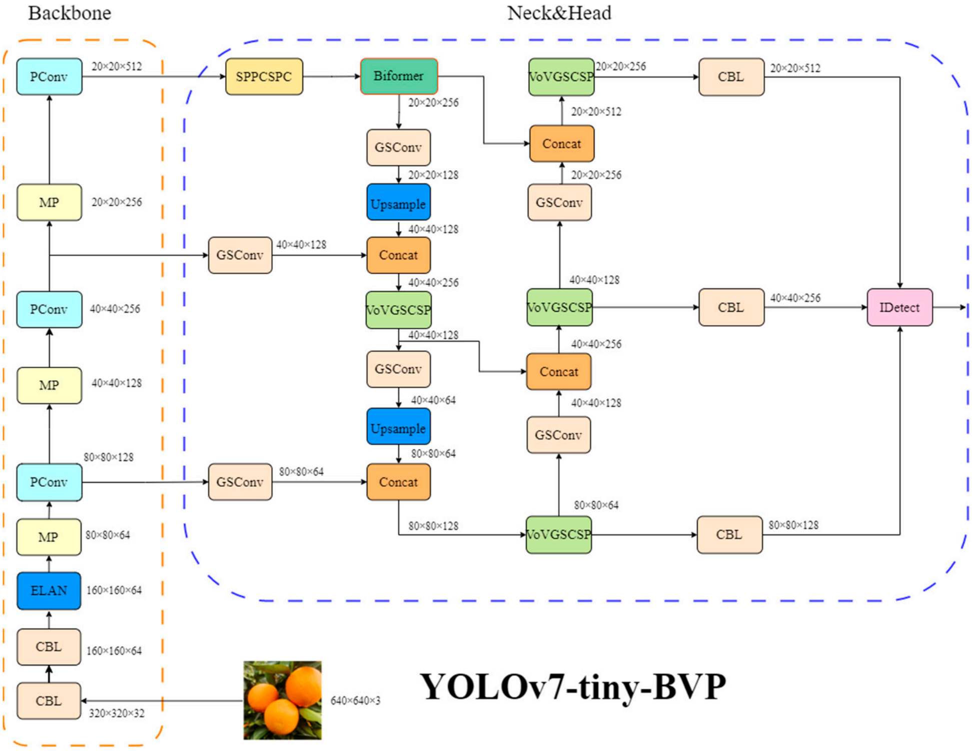


Figure 3. Diagram of the YOLOv7-tiny-BVP network.

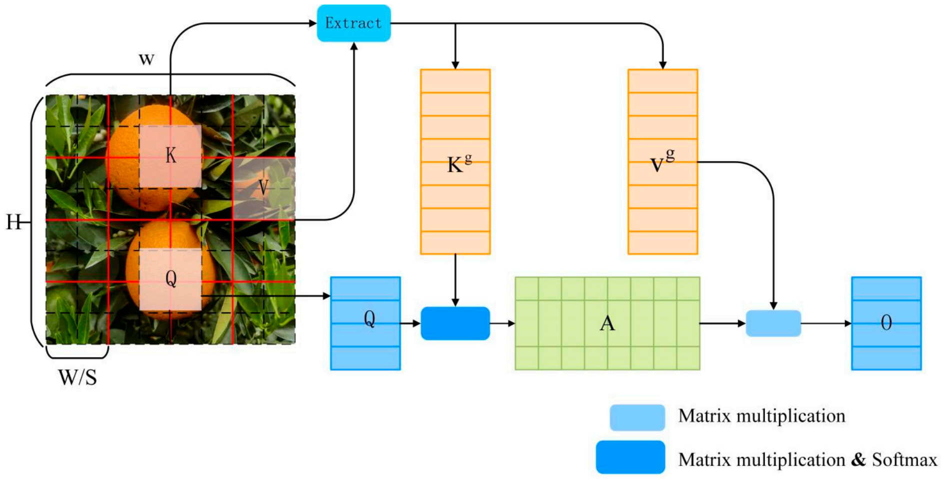


Figure 4. Diagram of the BiFormer attention mechanism.

The query vector Q is used to represent the content that the current position needs to focus on the attention mechanism. The weight in the attention allocation process is determined through calculating the similarity with the key vector K. The value vector V contains the actual feature representation, which is assigned different weights based on the similarity between Q and K to determine the importance of the different positions in the output. The relationships between them are shown in Equations (5)–(7):

$$\mathrm{K^g = gather(K, I^r)} \tag{5}$$

$$\mathrm{V^g = gather(V, I^r)} \tag{6}$$

$$\mathrm{O = Attention(Q, K^g, V^g) + LCE(V)} \tag{7}$$

where I^r in Equations (5) and (6) represents the index tensor, which is used to specify the position where the elements are collected from the input tensor. The gathering operation involves collecting the elements at corresponding positions from the input tensor based on the index; K^g and V^g are new key and value vectors obtained after the gathering operation. In Equation (7), attention represents the attention mechanism calculation, and LCE represents the cross-entropy loss function. O is the final output result that is calculated through the attention mechanism and combined with cross-entropy loss to obtain the final output, helping the model to better understand contextual information in detection tasks and to make accurate predictions.

3.2.2. VoVGSCSP Module Based on a Slim Neck

To further improve the model's ability to capture the positional information of citrus fruit features and reduce computational costs, the regular convolution in the neck network is replaced with GSConv, and the ELAN module is replaced with the VoVGSCSP module in this study to balance the model accuracy and speed [25].

In the convolutional neural network architecture, regular convolutional operations have a high accuracy, but the model is more complex and has a longer inference time. In contrast, although DWConv (depth-wise separable convolution) has a faster detection speed, it has a lower accuracy [26]. Therefore, the GSConv module was originally designed to reduce the model complexity while maintaining detection accuracy; the structure of the module is shown in Figure 5. GSConv infiltrates the semantic information generated by regular convolution into each part of DWConv, which fully utilizes the features of regular convolution and DWConv convolution, increasing both the speed of DWConv and the accuracy of regular convolution; thus, it achieves a better performance in citrus fruit detection tasks.

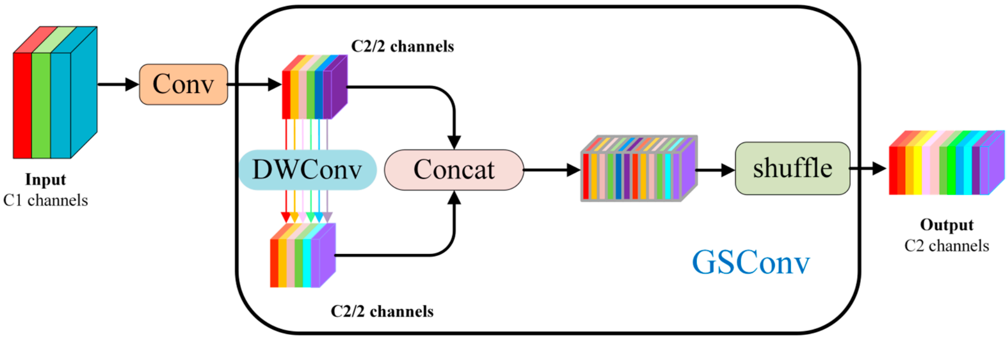


Figure 5. Diagram of the GSConv structure.

Although GSConv can significantly reduce the redundant information in the feature maps of citrus fruit detection models, it has limitations in further reducing inference time and maintaining accuracy. Therefore, the ELAN module is replaced with the VoVGSCSP module in the neck network section. The VoVGSCSP is formed by combining the modules in Figure 5 through a one-time aggregation method. The structure of VoVGSCSPC is shown in Figure 6. This approach can reduce the complexity of the computations and network structure and further minimize the memory footprint of the model, facilitating model deployment to edge devices with limited computational resources.

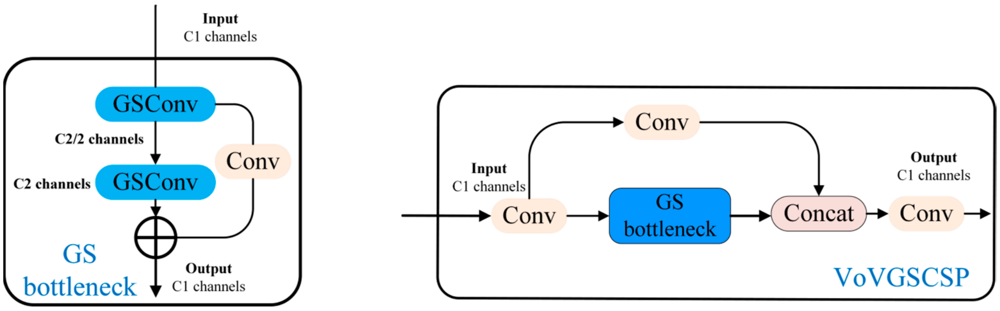


Figure 6. Structural diagram of the VoVGSCSPC.

3.2.3. PConv Module Based on FasterNet

The floating-point operation per second (FLOPS) is the calculation speed; the larger the value, the better the network performance. The current mainstream conventional convolution, group convolution, and depth-wise separable convolution (DWConv) methods all suffer from low FLOPS problems, which is mainly due to frequent access memory [27]. For floating-point of operations (FLOPs), the greater the number of computations, the smaller the value, which is generally used to measure the model complexity. The original YOLOv7-tiny backbone network uses large numbers of regular convolutions for feature extraction, with many parameters and a large computational effort. Although convolution kernels can be utilized to reduce the number of parameters and FLOPs, memory access increases as the network width increases, to compensate for the accuracy degradation [28]. Partial convolution (PConv) uses partial channels for spatial feature extraction, which can effectively reduce redundant computations and memory access. Therefore, to reduce the network complexity while maintaining feature extraction ability, this experiment replaces the regular convolution in the backbone network with PConv. The working principle of PConv is shown in Figure 7.

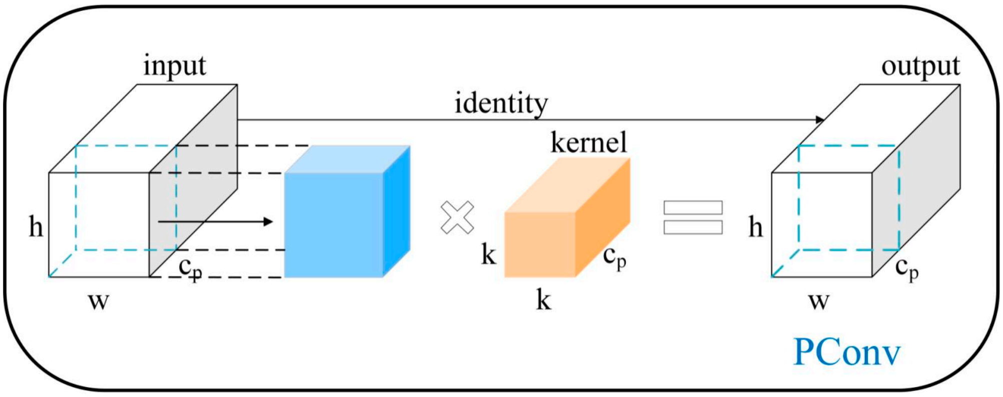


Figure 7. Diagram of the PConv structure.

In Figure 7, h and w represent the height and width of the feature map, respectively. c represents the total number of channels, cp represents the number of channels used, and the convolutional kernel size is k × k. In contrast to regular convolution, PConv requires only a portion of the input channels for spatial feature extraction, while keeping the remaining channels constant. For consecutive or regular memory access, the first or last consecutive channel is considered representative of the entire feature map for computation, and the input and feature maps are considered to have the same number of channels.

Without a loss of generality, the formula for calculating the FLOPs and memory access of PConv is as follows:

$$h \times w \times k^2 \times c_p^2 \tag{8}$$

$$h \times w \times 2c_p + k^2 \times c_p^2 \approx h \times w \times 2c_p \tag{9}$$

According to Equations (8) and (9), the FLOPs of PConv are only 1/16 of regular Conv, and the memory access is only 1/4 of the regular Conv.

4. Results and Discussion

4.1. Comparative Testing of Different Detection Models

To verify the ability of the new YOLOv7-tiny-BVP model to recognize citrus fruits, the current mainstream target detection networks in the YOLOv5 [29] series (YOLOv5s, YOLOv5x, YOLOv5n, YOLOv5l, and YOLOv5m) and YOLOv7 [22] series (YOLOv7, YOLOv7x, and YOLOv7-tiny) were selected for a comprehensive comparison in this experiment. During training, all the networks used the same dataset without loading the official default weights, and all the hyperparameters remained consistent. After training was completed, the weight of the network with the highest accuracy was selected. With the same test set, a total of 319 images were comprehensively evaluated. In this evaluation, the FPS metric was calculated as follows: when batch size = 1 is maintained, the result is calculated by dividing 1000 ms by the sum of image reasoning time, preprocessing time, and post-processing time. The results of the above target detection model after the completion of the test are shown in Table 2.

Table 2. Comparison of the results of different network models.

Models	Layers	Para	Size/mb	P/%	R/%	mAP@.5/%	FPS	F1
YOLOv5s	157	7,012,822	14.4	96.8%	88.5%	95.2%	76.80	0.93
YOLOv5x	322	86,173,414	173.1	97.0%	89.6%	95.6%	16.00	0.93
YOLOv5n	157	1,760,518	3.9	95.1%	87.4%	94.2%	91.90	0.91
YOLOv5l	267	46,108,278	92.8	96.6%	90.5%	95.6%	26.10	0.93
YOLOv5m	212	20,852,934	42.2	96.7%	89.0%	95.7%	42.57	0.93
YOLOv7	314	36,481,772	74.8	97.5%	91.6%	97.1%	34.96	0.95
YOLOv7x	362	70,782,444	142.1	97.6%	92.0%	96.8%	21.27	0.95
YOLOv7-tiny	172	6,007,596	12.3	97.0%	95.0%	98.0%	87.26	0.96
YOLOv7-tiny-BVP	257	**3,696,396**	**7.7**	**97.9%**	94.6%	97.8%	**89.28**	**0.97**

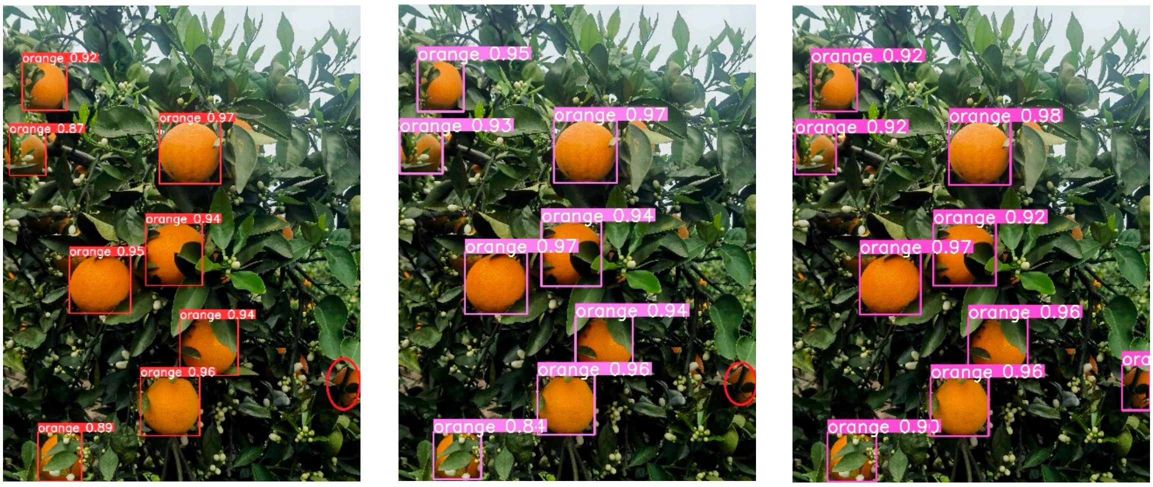
Note: Bold font is the optimal value for the model detection indicators; mAP@.5 average accuracy at IoU = 0.5.

Table 2 shows that the current YOLOv5 series network performs well in citrus fruit detection, but the overall performance is weaker than that of the YOLOv7 series network. Although YOLOv5n has a small number of parameters and a smaller proportion of models, it is still weaker than YOLOv7 series networks in terms of the other indicators. Although the P, R, and mAP indices of the YOLOv7 and YOLOv7x networks are higher than those of the YOLOv5 series, the overall model parameters are large, the model proportion is too large, and the FPS frame rate is low, making it difficult to deploy edge computing devices with insufficient GPU resources for detection.

Compared with YOLOv7-tiny, the overall parameters decreased by 38.47%, the model scale was reduced by 4.6 Mb, and P increased by 0.9%. Meanwhile, the FPS and F1 scores increased by 2.02 and 1%, respectively. Obviously, YOLOv7-tiny-BVP effectively reduces the computational load on the device, and still maintains a high performance.

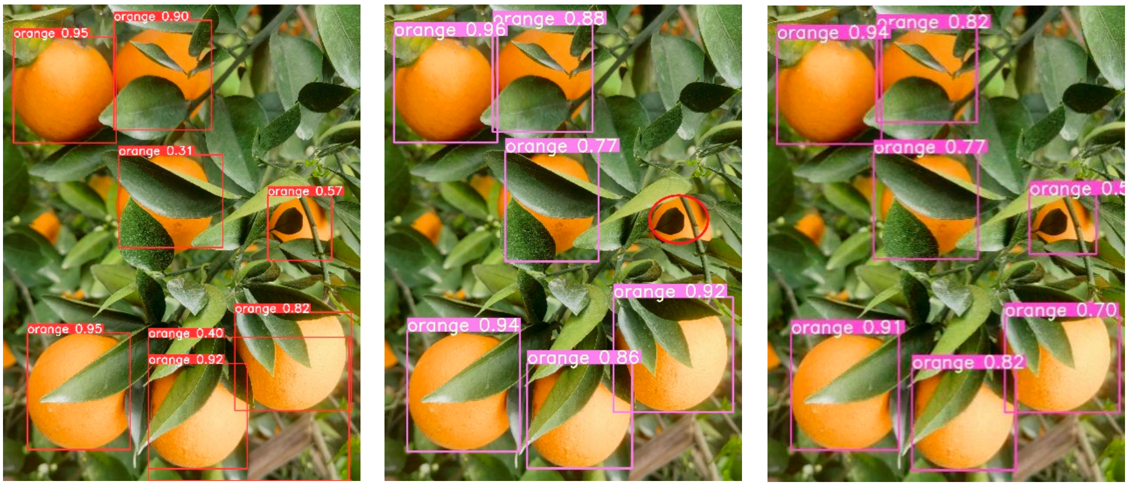
4.2. Comparative Experiments in Complex Scenarios

Under natural conditions, the orchard environment is more complex. The growth characteristics of fruit trees result in severe damage between the branches and leaves of the fruit. Moreover, fruits of different sizes and shapes commonly overlap each other. Additionally, there are many disturbing factors in the field environment, such as light and darkness, which make fruit identification and counting difficult [30]. Therefore, for the above two real-world scenarios, the YOLOv5s model, which has the most balanced performance among the YOLOv5 series; the YOLOv7-tiny; and the YOLOv7-tiny-BVP proposed in this paper were selected for comparative analysis. In Figures 8–11, the red and green circles indicate the network failing to detect under two different scenarios.



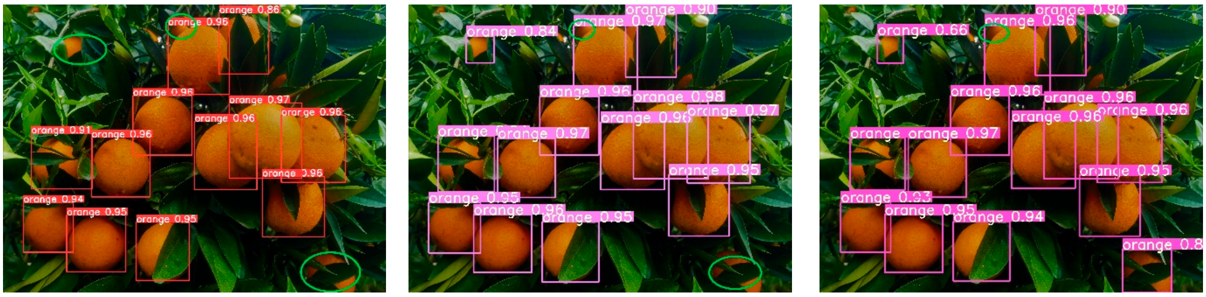
YOLOv5s YOLOv7-tiny YOLOv7-tiny-BVP

Figure 8. Mild occlusion recognition results.



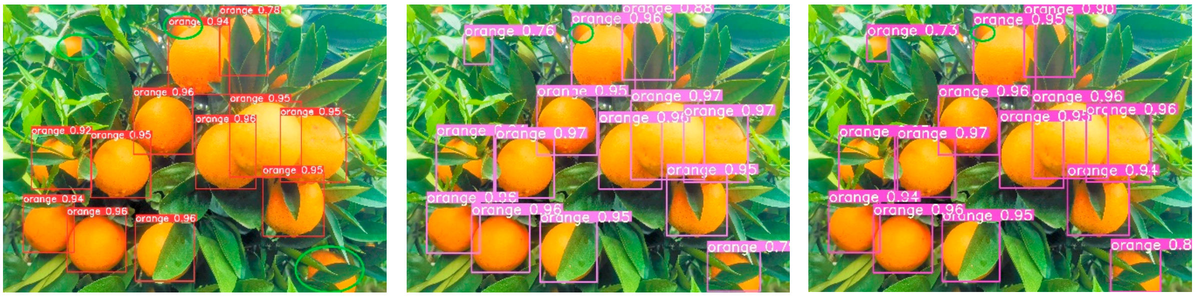
YOLOv5s YOLOv7-tiny YOLOv7-tiny-BVP

Figure 9. Heavy occlusion recognition results.



YOLOv5s YOLOv7-tiny YOLOv7-tiny-BVP

Figure 10. Recognition effect in a dark scene.



YOLOv5s YOLOv7-tiny YOLOv7-tiny-BVP

Figure 11. Recognition results in bright scenes.

4.2.1. Comparison of Recognition Results under Different Occlusion Scenes

Figure 8 shows that both YOLOv5s and YOLOv7-tiny can recognize more obvious targets in the slightly occluded scene, but the network is not sensitive to small targets at the edge of the image due to foliage occlusion, and both suffer from missed detections. YOLOv7-tiny-BVP can accurately detect the location of citrus fruit plants by introducing an attention mechanism, which increases the sensitivity of the network to the feature information around the citrus plants.

Figure 9 shows that in the case of heavy occlusion, overlapping fruits and leafy branch misalignment result in target feature loss. Although the improved network can still accurately recognize the target in the case of heavy occlusion, YOLOv7-tiny suffers from the omission of detection, and YOLOv5s can accurately recognize the fruits but suffers from misdiagnosis, incorrectly recognizing the two adhered fruits as one. In summary, the YOLOv7-tiny-BVP network proposed in this paper achieves a better performance in different occlusion scenarios.

4.2.2. Comparison of Recognition Results in Bright and Dark Scenes

Brightness also poses a considerable challenge for fruit identification and counting during actual field inspections. Figure 10 shows the actual detection results of the three network models for different light and dark scenes. As Figure 10 shows, in the case of dim light, YOLOv5s has a poor recognition performance compared to the other two networks, and three fruits were missed. YOLOv7 has two unrecognized fruits, while the recognition effect of YOLOv7-tiny-BVP is better overall than that of the other two networks. It is concluded that the color features of citrus fruit plants are masked in a dimly lit environment, the network cannot extract effective textural features, and there is a complication associated with overlapping leaves and fruits in the figure, leading to individual fruits not being detected.

In addition, in a well-lit environment, under the conditions of severe leaf and fruit overlap, the recognition ability of YOLOv7-tiny-BVP and YOLOv7-tiny proposed in this study significantly improved compared to that of the YOLOv5s network. Figure 11 shows that the recognition effect of YOLOv7-tiny is more accurate than that of dim light scenarios. It is concluded that under sufficient light conditions, the network is more likely to obtain semantic color information about fruits, resulting in more accurate fruit recognition than in dim light scenarios. Therefore, the YOLOv7-tiny-BVP network proposed in this paper is more suitable for fruit detection in complex environments, has a high accuracy, and fewer missed detections under different scenes of light and darkness.

4.3. *Ablation Experiment*

To validate the effectiveness of the three improvement strategies proposed in this article for the YOLOv7-tiny-BVP network, ablation experiments were conducted to observe the changes in various indicators on the same training set after 300 epochs of training. The training results are shown in Figure 12, where the model gradually starts to converge after 250 iterations, and there is no significant change between the three independent improvement schemes and the original network model YOLOv7-tiny in terms of precision, recall, and mAP@.5 in the training set. However, in terms of the mAP@.5:.95 indicator, the YOLOv7-tiny-BVP achieves better results when three improvement strategies are introduced simultaneously, rather than when the BiFormer attention mechanism, VoVGSCSP, or PConv are introduced each time alone. In addition, YOLOv7-tiny-BVP also has a lower localization (boss_loss) and confidence loss (obj_loss) than the other individual improvement strategies, indicating that the model overall performs better after introducing the three improvement strategies.

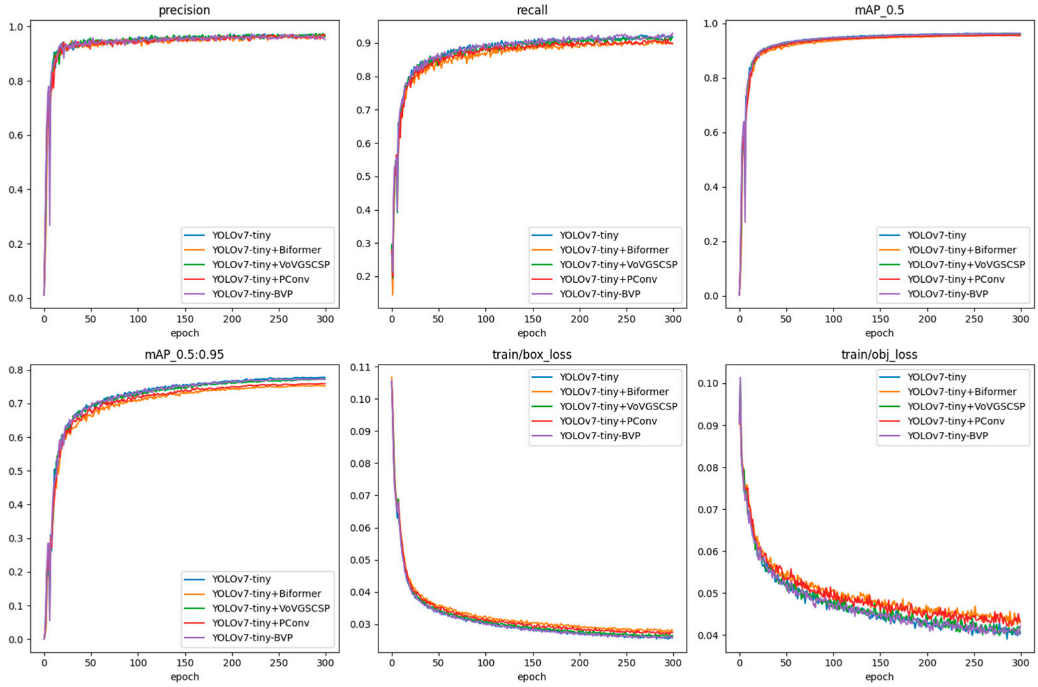


Figure 12. Ablation experiment.

After the training was completed, the weights with optimal accuracy were selected for a comprehensive model evaluation in the same validation set of 573 images; the evaluation results are shown in Table 3. In Table 3, time is the sum of the image preprocessing time, inference speed, and the NMS non extremely large suppression time; a smaller value indicates a better model performance. The analysis of various indicators in the table shows that the accuracy of the model improves by 1% after introducing the BiFormer attention mechanism. The analysis suggests that introducing an attention mechanism enhances the network's ability to extract features and improve detection accuracy. However, at the same time, the proportion of the model is relatively increased, which does not meet the need for lightweight materials. In addition, after introducing the VoVGSCSP module, the overall parameter quantity and the model size decrease slightly, while the accuracy improves by 1%. Moreover, the model's image processing time increases by 1.2 ms, resulting in a relatively poor real-time detection efficiency.

Moreover, after the ELAN module was replaced with the PConv module in the YOLOv7-tiny backbone network, the overall parameter quantity and the proportion of the model significantly decreases, while the accuracy decreases by 2% compared to those of the original network. Additionally, the analysis shows that the number of convolutions and the depth of the network decrease after the network is replaced with PConv, after which the number of parameters is reduced. Therefore, the network's feature extraction ability is weakened. In addition, after introducing the three improvement schemes simultaneously, the recall and mAP@.5 metrics of the model decrease by 1%, and the overall parameters of the model decrease significantly. The improved model has the highest accuracy and shortest image processing time, and the model is lighter, more accurate, more real-time, and easier to deploy.

Table 3. Comparison of ablation experiment indices.

Model	Layers	Para	Size/mb	P/%	R/%	mAP@.5/%	Time
YOLOv7-tiny	172	6,007,596	12.3	96.4	91.4	96.1	8.4 ms
YOLOv7-tiny + BiFormer	183	6,273,324	12.8	97.5	91.3	96.2	8.2 ms
YOLOv7-tiny + VoVGSCSP	267	5,620,492	11.6	97.1	88.7	95.5	9.6 ms
YOLOv7-tiny + PConv	151	3,817,772	7.8	94.3	91.7	95.5	8.2 ms
YOLOv7-tiny-BVP	257	**3,696,396**	**7.7**	**97.4**	89.7	95.5	**7.9 ms**

Note: Bold font is the optimal value for the model detection indicators; mAP@.5 average accuracy at IoU = 0.5.

4.4. Comparison of the Latest Methods for Fruit Testing

Table 4 shows the latest research results. Lai et al. [10] used the YOLOv7 network to detect pineapple fruits and classify pineapple fruits according to their ripeness. Ma et al. [18] detect apples under complex weather conditions by using the YOLOv7-tiny model and a public apple dataset. Compared to their model, our model has a significant improvement in P, F1 score, and mAP. Chen et al. [12] used the YOLOv7 network to detect the ripeness of citrus fruits; though the mAP of the model reached 97.29%, the model has too much computation and too many parameters. In addition, Zhang et al. [31] used UAV to collect citrus fruit images and employed the YOLOv7-tiny network to recognize the fruits; though the size of the model is only 3.96 MB, our model has a great improvement in both P and F1 scores. In summary, the model proposed in this study has a more comprehensive performance than the models compared above and is more suitable for citrus fruit detection.

Table 4. Latest research on fruit detection.

Authors	Crops	Model	P/%	mAP@.5/%	F1/%	Size/mb	Time/ms
Lai et al. [10]	Pineapple	YOLOv7	94.17	95.82	91.95	/	23.81
Ma et al. [18]	Apple	YOLOv7-tiny	80.1	80.4	76.8	5.06	/
Chen et al. [12]	Citrus	YOLOv7	94.25	97.29	93.81	24.26	69.38
Zhang et al. [31]	Citrus	YOLOv7-tiny	89.02	90.34	86.0	3.96	/
Ours	Citrus	YOLOv7-tiny	97.9	97.8	97.0	7.7	7.9

4.5. Discussion on Lightweight Networks

With the rapid development of mobile and embedded devices nowadays, more and more algorithms are deployed to mobile devices for people's convenience, but the aforementioned devices suffer from the problems of small memory, insufficient storage space, and limited computational power [32,33]. The lightweight network proposed in this study can both accurately identify citrus fruits and significantly reduce the number of network participants. As shown in Figure 13 the image is the side view of a citrus fruit tree, and the improved network maintains a high recognition rate for citrus fruits, which provides a new way to deploy the aforementioned mobile devices for fruit identification and yield estimation.

Although the network model proposed in this study is small and has a high accuracy, it still has some limitations. (1) The date and location of image collection is relatively homogeneous. (2) There is a lack of green or yellow-green citrus fruits in the images. (3) The proposed network only demonstrates the promise of a lightweight network and was not deployed to mobile devices for testing. In future research, we will focus more on image acquisition, including date, location, and type, to optimize the proposed lightweight network for the real-time detection of citrus fruits during the full-growth period.

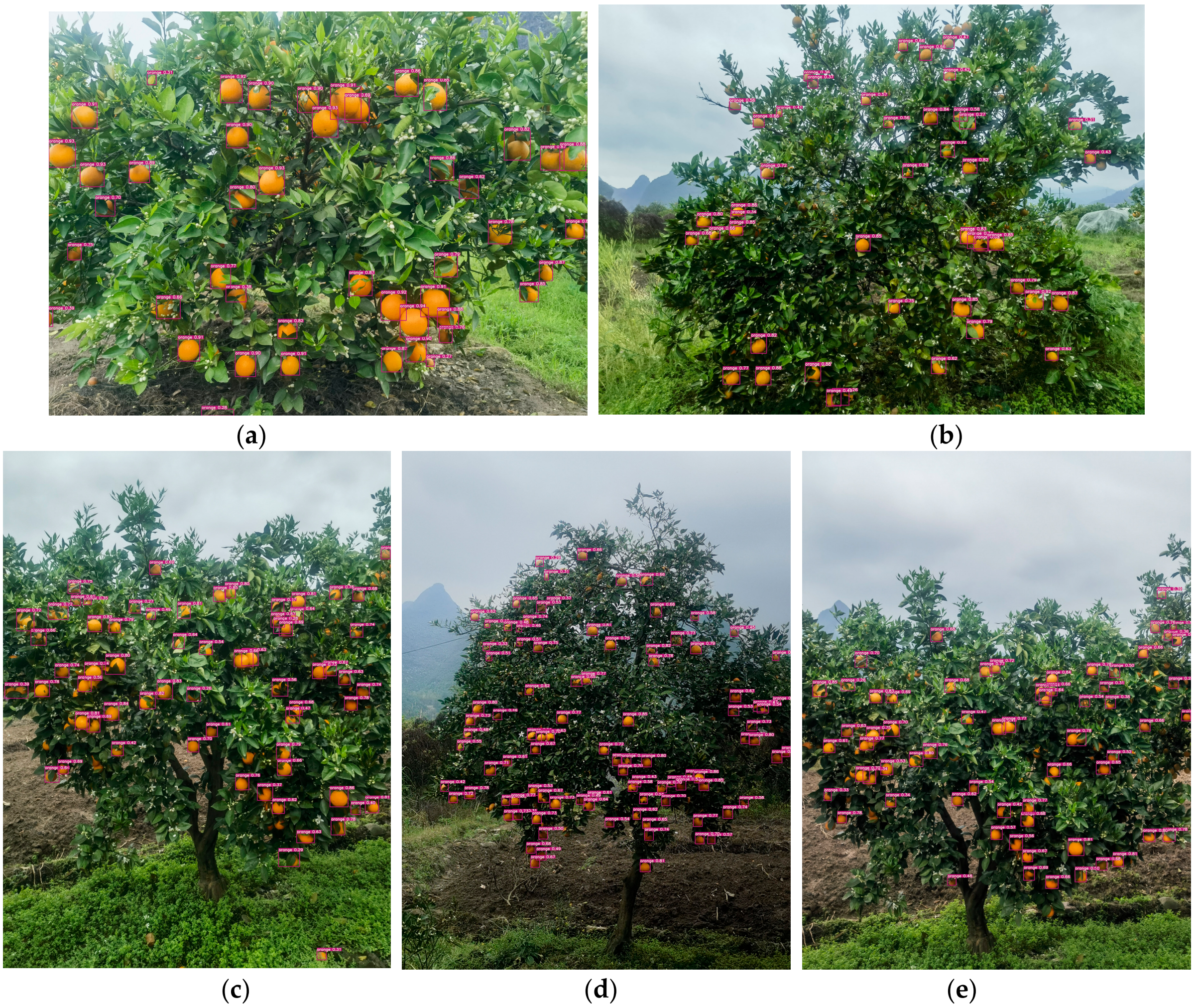


Figure 13. Fruit identification results of citrus fruit trees from a side view. (**a**–**e**) Results of different fruit tree tests.

5. Conclusions

Based on YOLOv7-tiny, a new network model, YOLOv7-tiny-BVP, is constructed in this study by introducing the BiFormer attention mechanism, the VoVGSCSPC module, and the PConv convolution. In this study, a total of eight different networks from the YOLOv5 and YOLOv7 series were selected for comparative testing with the same dataset. The results show that compared with YOLOv7-tiny, the total number of parameters and model scale of YOLOv7-tiny-BVP are reduced by 38.47% and 4.6 MB, respectively.

Meanwhile, the accuracy, FPS, and F1 were improved to different degrees, respectively, and the detection accuracy was no less than that of YOLOv7-tiny. Therefore, the proposed YOLOv7-tiny-BVP network model is not only lightweight, but also has the advantage of a high detection accuracy of the YOLOv7 model, which provides a new idea for the direction of the future research of lightweight models, and also provides the possibility of applying the lightweight target detection model in the field of agricultural yield estimation.

Author Contributions: Conceptualization, J.J., M.Z. and S.D.; Data curation, L.W., S.Y. and J.Y.; Formal analysis, J.J. and S.D.; Funding acquisition, J.J. and S.D.; Investigation, B.L., J.Y. and M.Z.; Methodology, S.D., J.J. and L.W.; Project administration, S.D.; Resources, M.Z. and J.J.; Software, M.Z., S.Y. and B.L.; Validation, M.Z. and J.Y.; Visualization, S.D. and J.J.; Writing—original draft, J.J., M.Z. and L.W.; Writing—Review and editing, J.J. and S.D. 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 (Nos. 42161028, 42061059), the Natural Science Foundation of Guangxi (No. 2023JJA150097), the Guilin Science and Technology Bureau Development Program (No. 2020010701, 20210226-2), and the "BaGui Scholars" program of the provincial government of Guangxi, the Guilin University of Technology Foundation (No. GUTQDJJ2019046).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Data will be made available on request.

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

Abbreviations

The following abbreviations are used in this manuscript:

ELAN	efficient layer aggregation network
PConv	partial convolution
FPS	frames per second
CBAM	convolutional block attention module
BiFPN	bidirectional feature pyramid network
SE	squeeze-and-excitation networks
DWConv	Depth-wise separable convolution

References

1. Zhang, F. Statistical Analysis of Fruit Production in China in 2020. *China Fruit News* **2021**, *38*, 29–39.
2. Wang, L.; Zhao, Y.; Xiong, Z.; Wang, S.; Li, Y.; Lan, Y. Fast and precise detection of litchi fruits for yield estimation based on the improved YOLOv5 model. *Front. Plant Sci.* **2022**, *13*, 965425. [CrossRef]
3. Sozzi, M.; Cantalamessa, S.; Cogato, A.; Kayad, A.; Marinello, F. Automatic Bunch Detection in White Grape Varieties Using YOLOv3, YOLOv4, and YOLOv5 Deep Learning Algorithms. *Agronomy* **2022**, *12*, 319. [CrossRef]
4. Wang, D.; He, D. 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]
5. Cardellicchio, A.; Solimani, F.; Dimauro, G.; Petrozza, A.; Summerer, S.; Cellini, F. Detection of tomato plant phenotyping traits using YOLOv5-based single stage detectors. *Comput. Electron. Agric.* **2023**, *207*, 107757. [CrossRef]
6. Liu, Y.; Ren, H.; Zhang, Z.; Men, F.; Zhang, P.; Wu, D.; Feng, R. Research on multi-cluster green persimmon detection method based on improved Faster RCNN. *Front. Plant Sci.* **2023**, *14*, 1177114. [CrossRef]
7. Wang, C.; Wang, Y.; Liu, S.; Lin, G.; He, P.; Zhang, Z.; Zhou, Y. Study on Pear Flowers Detection Performance of YOLO-PEFL Model Trained with Synthetic Target Images. *Front. Plant Sci.* **2022**, *13*, 911473. [CrossRef]
8. Zhou, J.; Zhang, Y.; Wang, J. RDE-YOLOv7: An Improved Model Based on YOLOv7 for Better Performance in Detecting Dragon Fruits. *Agronomy* **2023**, *13*, 1042. [CrossRef]
9. Zhou, J.; Zhang, Y.; Wang, J. A Dragon Fruit Picking Detection Method Based on YOLOv7 and PSP-Ellipse. *Sensors* **2023**, *23*, 3803. [CrossRef]
10. Lai, Y.; Ma, R.; Chen, Y.; Wan, T.; Jiao, R.; He, H. A Pineapple Target Detection Method in a Field Environment Based on Improved YOLOv7. *Appl. Sci.* **2023**, *13*, 2691. [CrossRef]
11. Bi, S.; Gao, F.; Chen, J.; Zhang, L. Detection Method of Citrus Based on Deep Convolution Neural Network. *Trans. Chin. Soc. Agric. Mach.* **2019**, *50*, 181–186.
12. Chen, J.; Liu, H.; Zhang, Y.; Zhang, D.; Ouyang, H.; Chen, X. A Multiscale Lightweight and Efficient Model Based on YOLOv7: Applied to Citrus Orchard. *Plants* **2022**, *11*, 3260. [CrossRef]
13. Zhao, C.; Liang, X.; Yu, H.; Wang, H.; Fan, S.; Li, B. Automatic Identification and Counting Method of Caged Hens and Eggs Based on Improved YOLOv7. *Trans. Chin. Soc. Agric. Mach.* **2023**, *54*, 300–312.
14. Xiong, J.; Zheng, Z.; Liang, J.; Zhong, Z.; Liu, B.; Sun, B. Citrus Detection Method in Night Environment Based on Improved YOLO v3 Network. *Trans. Chin. Soc. Agric. Mach.* **2020**, *51*, 199–206.
15. Huang, H.; Hu, J.; Li, Z.; Wei, Z.; Liu, S. Design of citrus fruit intelligent recognition system based on edge computing. *J. Hunan Agric. Univ. (Nat. Sci.)* **2021**, *47*, 727–732.
16. Huang, T.; Huang, H.; Li, Z.; Lv, S.; Xue, X.; Dai, q.; Wen, W. Citrus fruit recognition method based on the improved model of YOLOv5. *J. Huazhong Agric. Univ.* **2022**, *41*, 170–177. [CrossRef]
17. Wang, Z.; Jin, L.; Wang, S.; Xu, H. Apple stem/calyx real-time recognition using YOLO-v5 algorithm for fruit automatic loading system. *Postharvest Biol. Technol.* **2022**, *185*, 111808. [CrossRef]
18. Ma, L.; Zhao, L.; Wang, Z.; Zhang, J.; Chen, G. Detection and Counting of Small Target Apples under Complicated Environments by Using Improved YOLOv7-tiny. *Agronomy* **2023**, *13*, 1419. [CrossRef]

19. Wang, F.; Jiang, J.; Chen, Y.; Sun, Z.; Tang, Y.; Lai, Q.; Zhu, H. Rapid detection of Yunnan Xiaomila based on lightweight YOLOv7 algorithm. *Front. Plant Sci.* **2023**, *14*, 1200144. [CrossRef]
20. Mekhalfi, M.L.; Nicolo, C.; Ianniello, I.; Calamita, F.; Goller, R.; Barazzuol, M.; Melgani, F. Vision System for Automatic On-Tree Kiwifruit Counting and Yield Estimation. *Sensors* **2020**, *20*, 4214. [CrossRef]
21. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
22. Wu, C.; Ye, M.; Zhang, J.; Ma, Y. YOLO-LWNet: A Lightweight Road Damage Object Detection Network for Mobile Terminal Devices. *Sensors* **2023**, *23*, 3268. [CrossRef] [PubMed]
23. Yang, Z.; Feng, H.; Ruan, Y.; Weng, X. Tea Tree Pest Detection Algorithm Based on Improved Yolov7-Tiny. *Agriculture* **2023**, *13*, 1031. [CrossRef]
24. Zhu, L.; Wang, X.; Ke, Z.; Zhang, W.; Lau, R.W. BiFormer: Vision Transformer with Bi-Level Routing Attention. *arXiv* **2023**, arXiv:2303.08810. [CrossRef]
25. Li, H.; Li, J.; Wei, H.; Liu, Z.; Zhan, Z.; Ren, Q. Slim-neck by GSConv: A better design paradigm of detector architectures for autonomous vehicles. *arXiv* **2022**, arXiv:2206.02424. [CrossRef]
26. Dai, Y.; Li, C.; Su, X.; Liu, H.; Li, J. Multi-Scale Depthwise Separable Convolution for Semantic Segmentation in Street-Road Scenes. *Remote Sens.* **2023**, *15*, 2649. [CrossRef]
27. Liu, C.; Wang, T.; Dong, S.; Zhang, Q.; Yang, Z.; Gao, F. Hybrid Convolutional Network Combining 3D Depthwise Separable Convolution and Receptive Field Control for Hyperspectral Image Classification. *Electronics* **2022**, *11*, 3992. [CrossRef]
28. Chen, J.; Kao, S.; He, H.; Zhuo, W.; Wen, S.; Lee, C.H.; Chan, S.H.G. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. *arXiv* **2023**, arXiv:2303.03667. [CrossRef]
29. Zhu, X.; Liu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios. *arXiv* **2021**, arXiv:2108.11539. [CrossRef]
30. Dorj, U.O.; Lee, M.; Yun, S. A yield estimation in citrus orchards via fruit detection and counting using image processing. *Comput. Electron. Agric.* **2017**, *140*, 103–112. [CrossRef]
31. Zhang, Y.; Fang, X.; Guo, J.; Wang, L.; Tian, H.; Yan, K.; Lan, Y. CURI-YOLOv7: A Lightweight YOLOv7tiny Target Detector for Citrus Trees from UAV Remote Sensing Imagery Based on Embedded Device. *Remote Sens.* **2023**, *15*, 4647. [CrossRef]
32. Yin, H. Research on Fall Detection Algorithm and Algorithm Deployment in Embedded Platform. Master's Thesis, University of Electronic Science and Technology of China, Chengdu, China, 2023.
33. Zhu, H. Research on Lightweight and Mobile Deployment Method of Road Target Detection Algorithm Based on Deep Learning. Master's Thesis, Inner Mongolia Agricultural University, Hohhot, China, 2022.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.