Lightweight detection method for lotus seedpod in natural environment

Tao Tang[1], Xu Wang[1], Zenghong Ma[1,2], Weiwei Hong[1,3], Gaohong Yu[1,2], Bingliang Ye[1,2*]

(1. *Faculty of Mechanical Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China*;
2. *Key Laboratory of Transplanting Equipment and Technology of Zhejiang Province, Hangzhou 310018, China*;
3. *Special Equipment Institute, Hangzhou Vocational & Technical College, Hangzhou 310018, China*)

Abstract: In order to solve the problems of the current target detection algorithms, such as poor discrimination of occluded targets, multiple parameters, complex networks, large amounts of computation, and not conducive to the deployment of mobile terminals, a lightweight lotus seedpod detection method based on YOLOv5s model was proposed in this study. First, the dataset was augmented by using a combination of offline and online augmentation, which improved the adaptability and robustness of the model in complex environments. Then, a lightweight Ghost convolution module was introduced to replace the original convolution, and a lightweight bidirectional feature pyramid network was designed, which could enhance the feature extraction and fusion capability of the network and reduce the amount of calculation and model size; On this basis, the combination of WIoU loss function and Mish activation function was adopted to improve the accuracy of feature extraction. Finally, the knowledge distillation training strategy was used to ensure the proposed lightweight model has the learning ability of a complex network model, improving the recall and precision of model detection. The results of the ablation study show that the proposed method effectively improves the detection performance of the YOLOv5s model for lotus seedpods. The mean average precision of the improved model was 89.7%, compared with the original YOLOv5s model increased by 2.8%, and the parameters and FLOPs were reduced by 2.36M and 7.3G, respectively. Compared with other detection algorithm models, the proposed algorithm model has the advantages of less computation, smaller model size, and higher detection precision. Therefore, the proposed improved optimization method based on the YOLOv5s model can effectively detect lotus seedpods, which provides theoretical research and technical support for intelligent picking of lotus seedpods in the actual operating environment.

Keywords: lotus seedpod detection, deep Learning, data augmentation, lightweight, knowledge distillation, natural environment

DOI: 10.25165/j.ijabe.20231606.8281

Citation: Tang T, Wang X, Ma Z H, Hong W W, Yu G H, Ye B L. Lightweight detection method for lotus seedpod in natural environment. Int J Agric & Biol Eng, 2023; 16(6): 197–206.

1 Introduction

As a traditional Chinese selectively harvested cash crop, the lotus seedpods are mainly harvested for the purpose of harvesting lotus seeds, which is characterized by strong regional production, extensive cultivation area, long life cycle, rich nutritional value, and high economic value[1,2]. In recent years, the planting area and yield of lotus in China have reached 613 thousand hm^2 and 15.48 million t, respectively. China has become a major producer and exporter of lotus seeds[3]. However, with the shortage of labor in the seasonal agricultural industry and soaring labor costs, the development of the lotus industry is facing the constraining problem of difficulties in harvesting lotus seedpods[4-6]. The current manual harvesting of lotus seedpods requires shuttle operations in the lotus pond mud, which has problems of difficulty in terms of labor demand and harvesting, low harvesting quality and efficiency, and personal safety hazards, so it is very urgent and important to design a high-quality lotus seedpods selective harvesting robot to realize intelligent harvesting of lotus seedpods. The natural growth environment of lotus seedpods is complex, with problems such as uneven lighting, branch and leaf shading, inconspicuous color differences with lotus leaves, etc.[7,8] Moreover, due to the limited computing power resources of the embedded platform on which the harvesting robot is equipped, the complex model cannot meet the real-time requirements of the task and is also difficult to deploy. Therefore, it is the key to accurately and quickly detect and identify the lotus seedpods for achieving intelligent harvesting[9,10], and it is also of great scientific importance and application value to study the lightweight detection algorithm of lotus seedpods in natural environment.

In recent years, with the rapid development of deep learning techniques[11], convolutional neural networks can extract multilevel features by unsupervised or semi-supervised feature learning and have stronger generalization capability than manually extracted features[12-14], so more and more deep learning algorithms are used for target recognition and detection tasks of agricultural robots in unstructured environments[15-19]. Wang et al.[20] proposed the CA-ENet model for identifying different apple diseases, which integrated a coordinate attention block in the EfficientNet-B4 network, used deeply differentiable convolution in the convolution module, and introduced the h-swish activation function. The results show that the proposed method can achieve competitive performance on the apple disease identification task. Kang et al.[21] proposed the DaSNet-v2 model for ripe apple detection, which combined instance segmentation branches and semantic

Received date: 2023-04-04 **Accepted date:** 2023-09-07

Biographies: Tao Tang, PhD candidate, research interest: intelligent agricultural equipment, Email: ttzstu@163.com; **Xu Wang**, MS, research interest: intelligent agricultural equipment, Email: 1755582156@qq.com; **Zenghong Ma**, PhD, Associate Professor, research interest: intelligent agricultural equipment, Email: mzh2018@zstu.edu.cn; **Weiwei Hong**, PhD candidate, research interest: intelligent agricultural equipment, Email: will_hong@yeah.net; **Gaohong Yu**, PhD, Professor, research interest: intelligent agricultural equipment, Email: yugh@zstu.edu.cn.

***Corresponding author: Bingliang Ye**, PhD, Professor, research interest: intelligent agricultural equipment. Faculty of Mechanical Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China. Tel: +86-13336060776, Email: zist_ybl@zstu.edu.cn.

segmentation branches into the architecture of a first-level detection network. Feature pyramid networks and Atrous spatial pyramid pooling were used to improve the performance of fruit and branch detection and segmentation. Tian et al.[22] proposed an improved YOLOv3 model for the detection of apples at different growth stages using DenseNet instead of Darknet53, which had better performance in the detection under overlapping and shading conditions. Ma et al.[23] proposed an improved YOLOv5 network model based on the coordinate attention (CA) module, which had better detection performance for detecting overripe lotus seedpods in different scenarios.

Although the above studies have achieved good detection results in the field of target detection, the number of network layers has been deepened and the number of model parameters has been increased due to the powerful feature extraction capability and robustness of convolutional neural networks. Therefore, many researches have also started to focus on the lightweight of target detection algorithms. Bhagat et al.[24] proposed a lightweight WheatNet-Lite architecture for wheat spikes detection, which integrated Mixed Depthwise Conv (MDWConv), Modified Spatial Pyramidal Polling (MSPP), and Depthwise Convolution (DWConv), with 54.2 M fewer network parameters compared with YOLOv3. Zha et al.[25] proposed YOLOv4_MF model for pest detection, using MobileNetV2 as a feature extraction block to reduce model parameters and focus loss instead of cross-entropy loss, and designing an improved feature fusion structure, finally the mean average precision (mAP) of the model was 4.24% higher than YOLOv4, while the volume was reduced to 1/6 of YOLOv4. Cui et al.[26] proposed a YOLOv4-Tiny model for pine cone detection, using LESNet as the backbone to extract pine cone features and a feature pyramid network with SE attention to fuse multi-scale information, and the average precision of the improved model was improved by 56.4% over the original, and the parameters and computation were compressed to 12.22% and 17.35% of the original network, respectively.

The above researches show that the lightweight target detection algorithms can effectively reduce the scale of the model and improve the detection precision. Therefore, in consideration of no study on the lightweight model of lotus seedpod detection, a lightweight lotus seedpod detection algorithm model based on YOLOv5s was constructed. YOLOv5s can achieve faster inference speed while maintaining high precision. By improving the model, the model params and model size were reduced, which improved the real-time performance of detection while the model precision was maintained.

2 Materials and methods

2.1 Image acquisition

Dataset creation is a key step in deep learning algorithms for target detection. In this study, lotus seedpods were taken as the research object, and the image data were collected at a thousand-mu lotus pond planting base in Canal Street, Linping District, Hangzhou, Zhejiang Province, China, which were photographed from August to October 2022. In order to fully consider the complexity of natural scenes, image acquisition was carried out from different illumination, angles, and distances. It is convenient for the target detection algorithm to better learn the detailed features of the labeled lotus seedpods target, so as to improve the recognition accuracy of the overall model. Figure 1 shows the schematic diagram of the experimental image acquisition process. The image acquisition was carried out by aerial view photography, and images were acquired at three time points: morning, afternoon, and evening, as well as under three occlusion conditions: no occlusion, slight occlusion, and severe occlusion. The acquisition device was an MV-CA013-20GC type Haikon camera, and a total of 1853 high-definition lotus seedpods images were finally acquired, with the image format of BMP and the resolution of 1280 pixels×1024 pixels for each image.

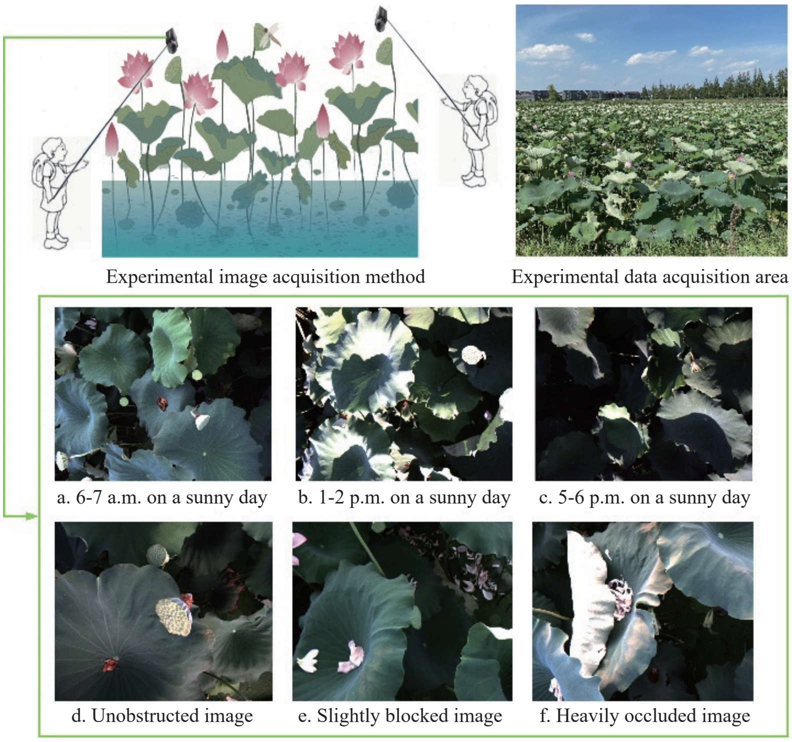

Experimental image acquisition method
Experimental data acquisition area
a. 6-7 a.m. on a sunny day
b. 1-2 p.m. on a sunny day
c. 5-6 p.m. on a sunny day
d. Unobstructed image
e. Slightly blocked image
f. Heavily occluded image

Figure 1 Schematic diagram of the acquisition process of experimental images

2.2 Image dataset construction

In order to improve the quality of the experimental dataset, different data augmentation methods as well as combined augmentation methods were used to train and test the augmented lotus seedpods dataset[27]. The original dataset was divided into training, validation, and test sets in the ratio of 7:1:2. The augmented dataset was divided into training set and validation set in the ratio of 8:2. The results of the division of the datasets are listed in Table 1.

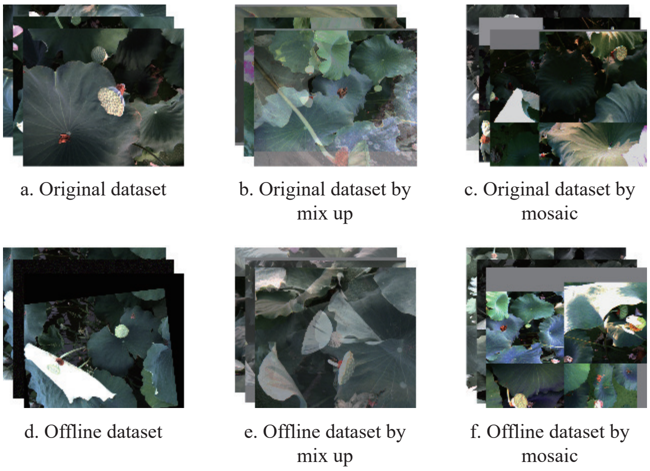
Table 1 Dataset partitioning

Datasets	Training set	Validation set	Test set	Total
Original	1297	185	371	1853
Original+Mix up	2780	926		3706
Original+Mosaic	2780	926		3706
Offline	2780	926		3706
Offline+Mix up	5930	1482		7412
Offline+Mosaic	5930	1482		7412

Prior to the training of the network model, the richness of the experimental dataset was enhanced by employing Mix up augmentation, Mosaic augmentation, and offline augmentation (adding noise, changing luminance, simulating occlusion, and performing affine transformation), as shown in Figure 2. The robustness of the model was improved by enhancing the image features and preventing overfitting.

2.3 Overview of the method

Compared with the two-stage detection network RCNN target detection network, YOLO network can significantly improve the operation speed of the model while keeping the detection accuracy basically unchanged[28]. YOLO, one of the available fastest target detection models, can directly obtain the class and the estimated probability of the target. In this study, considering the detection accuracy and lightweight requirements of the network, YOLOv5s was used as the basic framework of the lotus seedpod detection model. First, images of lotus seedpods in lotus ponds were acquired by cameras, and different data augmentation methods were used to construct the detection dataset of lotus seedpods. Then, a lightweight lotus seedpod detection model based on YOLOv5s was proposed, and the model was improved by modifying the backbone network and neck network, so as to improve the lotus seedpod feature extraction while also making lightweight improvements to the model. Finally, the knowledge distillation strategy was used to train the model and further improve the network accuracy. An overview of the method's technical route is shown in Figure 3.



a. Original dataset

b. Original dataset by mix up

c. Original dataset by mosaic

d. Offline dataset

e. Offline dataset by mix up

f. Offline dataset by mosaic

Figure 2 Dataset augmentation sample images

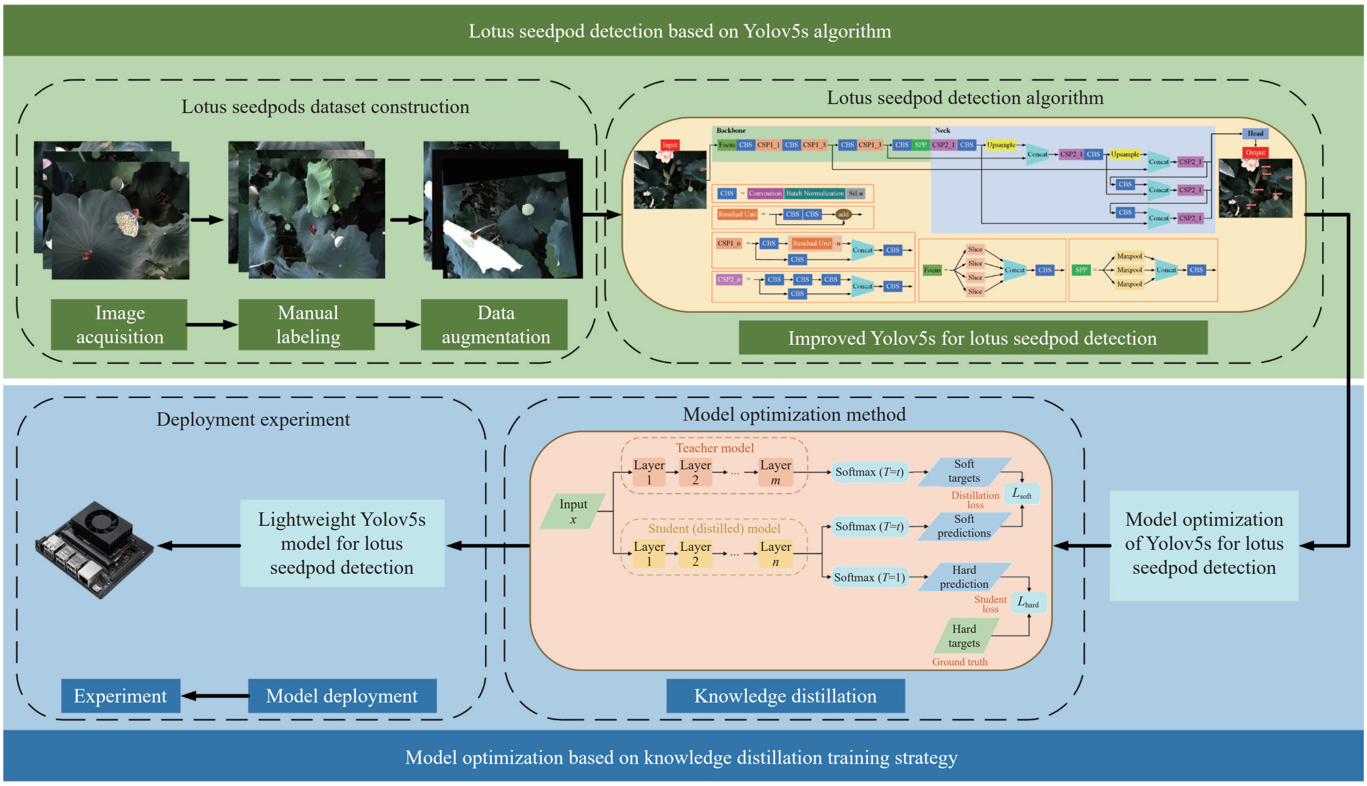

Lotus seedpod detection based on Yolov5s algorithm
Lotus seedpods dataset construction
Image acquisition
Manual labeling
Data augmentation
Lotus seedpod detection algorithm
Backbone
Neck
Head
Input
Output
Improved Yolov5s for lotus seedpod detection
Deployment experiment
Lightweight Yolov5s model for lotus seedpod detection
Experiment
Model deployment
Model optimization method
Teacher model
Layer 1
Layer 2
Layer m
Student (distilled) model
Layer n
Input x
Softmax (T=t)
Softmax (T=1)
Soft targets
Soft predictions
Distillation loss
L_{soft}
Hard prediction
Student loss
L_{hard}
Hard targets
Ground truth
Knowledge distillation
Model optimization of Yolov5s for lotus seedpod detection
Model optimization based on knowledge distillation training strategy

Figure 3 Overall technical route of the proposed lotus seedpod detection algorithm

2.4 Improvement of the YOLOv5s model

The model was improved with YOLOv5s as the base network. The improvement processes are as follows: 1) GhostNet was used to replace the DarkNet backbone; 2) Bi-directional Feature Pyramid Network (BiFPN) structure was introduced; 3) Group Shuffle Convolution (GSConv) was used to replace Conv at the neck layer; 4) VoV-GSCSP was used to replace C3 at the neck layer. The final network structure is shown in Figure 4.

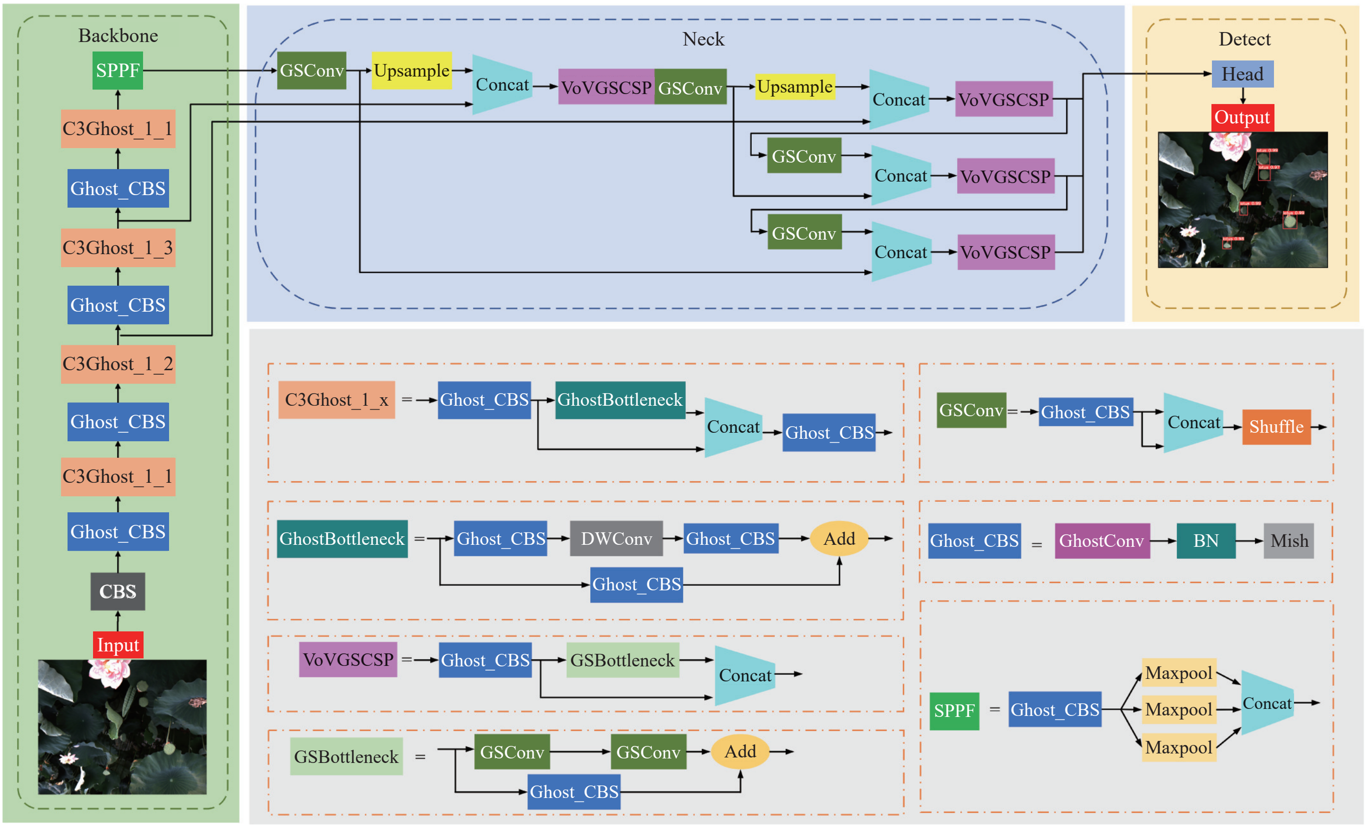
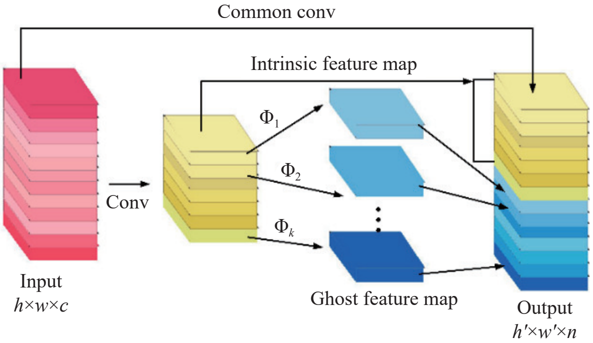

Backbone
SPPF
C3Ghost_1_1
Ghost_CBS
C3Ghost_1_3
Ghost_CBS
C3Ghost_1_2
Ghost_CBS
C3Ghost_1_1
Ghost_CBS
CBS
Input
Neck
GSConv
Upsample
Concat
VoVGSCSP
GSConv
Upsample
Concat
VoVGSCSP
GSConv
Concat
VoVGSCSP
GSConv
Concat
VoVGSCSP
Detect
Head
Output
C3Ghost_1_x = Ghost_CBS GhostBottleneck Concat Ghost_CBS
GSConv = Ghost_CBS Concat Shuffle
GhostBottleneck = Ghost_CBS DWConv Ghost_CBS Add Ghost_CBS
Ghost_CBS = GhostConv BN Mish
VoVGSCSP = Ghost_CBS GSBottleneck Concat
SPPF = Ghost_CBS Maxpool Maxpool Maxpool Concat
GSBottleneck = GSConv GSConv Add Ghost_CBS

Figure 4 Structure of the improved Yolov5s model

2.4.1 Improvement of the backbone network

The YOLOv5s network model contains more layers of CBL convolutional blocks, consisting of convolutional layers, batch normalization layers, and activation layers[29]. However, due to the limitation of storage space and computational power resources on mobile, the computational volume and size of the model need to be further reduced to enable subsequent deployment on mobile. Therefore, Ghost convolution was proposed to replace the normal convolution in the YOLOv5s network model, and Figure 5 shows the structure of Ghost convolution.



Note: h, w, and c are the height, width, and number of channels of the input image, respectively; h', w', and n are the height, width, and number of channels of the output image, respectively; Φ_1, Φ_2, and Φ_k are linear transformations.

Figure 5 Structure diagram of Ghost module

Ghost convolution can extract more features with fewer parameters compared to ordinary convolution. As shown in Figure 5, the Ghost module first performs the regular convolution operation on the input feature layer to generate a part of the real feature layer. Then the Ghost feature layer is obtained by performing a linear transformation on each channel of the real feature layer. Finally, the real feature layer and Ghost feature layer are connected to obtain a complete output feature layer. Suppose the input feature map is h×w×c, the output feature map is h'×w'×n, the convolution kernel size is k×k, and the input feature layer is divided into s parts. The floating-point operation for ordinary convolution, the floating-point operation of Ghost convolution, and the theoretical acceleration ratio of Ghost convolution instead of ordinary convolution operation are defined as follows, respectively.

$$\text{Flop1}=n\times h'\times w'\times c\times k\times k \tag{1}$$

$$\text{Flop2}=\frac{n}{s}\times h'\times w'\times c\times k\times k+(s-1)\times h'\times w'\times\frac{n}{s}\times k\times k \tag{2}$$

$$r_s=\frac{\text{Flop1}}{\text{Flop2}}=\frac{n\times h'\times w'\times c\times k\times k}{\frac{n}{s}\times h'\times w'\times c\times k\times k+(s-1)\times h'\times w'\times\frac{n}{s}\times k\times k}=\frac{c\times k\times k}{\frac{1}{s}\times c\times k\times k+\frac{s-1}{s}\times k\times k}\approx\frac{s\times c}{s+c-1}\approx s \tag{3}$$

It can be found that compared with the ordinary convolutional operation, Ghost convolutional operation has lower computational consumption. The original operation of generating feature maps using all convolutional kernels is changed to retaining only a small number of convolutional kernels, and other parts of convolutional operations are replaced by Ghost convolutional operation, which can achieve a significant reduction in the amount of computation and time required to generate feature maps.

2.4.2 Improvement of the neck network

In order to ensure the precision and speed of lotus seedpod detection during the automatic picking process, a lightweight bidirectional feature pyramid network was designed, which consists of BiFPN network structure, GSConv, and VoV-GSCSP module. Since lotus seedpods grow in complex environments, there are often detection errors and omission problems when they are obscured or small. To solve these problems for improving the detection performance of the model for obscured or small targets, a BiFPN network structure based on PANet was introduced, as shown in Figure 6. Although PANet can effectively fuse different feature layers, it is essentially a simple summation of different features. In contrast, BiFPN adopts a cross-scale connectivity approach to fuse features in the feature extraction network directly with features in the bottom-up path relative to size by adding an extra edge, so that the network retains more shallow semantic information without losing too much relatively deep semantic information, thus alleviating the inaccurate recognition problem of lotus seedpods caused by occlusion or smaller size.

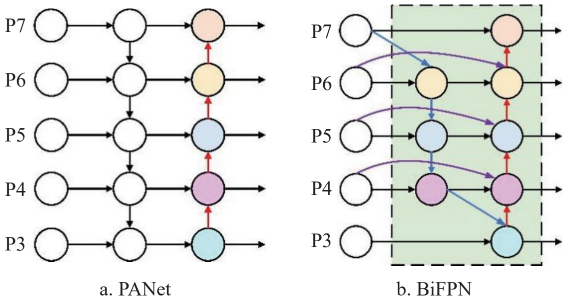


Figure 6 Structure of the PANet and BiFPN

The introduction of BiFPN module has a certain improvement in detection accuracy, but it introduces a certain number of parameters, which is not optimal for application to the accurate and fast identification of detection models for lotus seedpods. In order to enhance the network feature extraction and fusion capability while the params and FLOPs of the model can be reduced, the GSConv and VoV-GSCSP module were proposed, as shown in Figure 7. In the basic structure of lightweight networks, the depth-wise separable convolution (DSC) can effectively reduce the params and FLOPs of the model, but the channel information of the input image is separated by DSC during the computation, which leads to the feature extraction and fusion capability of DSC is much lower than that of standard convolution (SC). To make up for the shortcoming, the GSConv module consisting of the Conv module, the DSConv module, the Concat module, and the shuffle module was proposed, which effectively utilizes the computational power of DSC while enabling the detection precision of DSC to reach the standard of SC. With the introduction of the GSConv module, the params and FLOPs effort is reduced, but to enhance the expressiveness of the algorithmic model, the cross-stage partial network module VoV-GSCSP designed by using the one-shot aggregation method was proposed, which consists of the Conv module, the GS bottleneck module, and the Concat module. The module adds a new jump connection to the GS bottleneck so that the two branches perform separate convolutions without sharing weights. It also splits the number of channels using the split channel method so that the number of channels is propagated through different network paths, thus reducing the computational effort of the model and the complexity of the network structure while ensuring the accuracy of the propagated channel information.

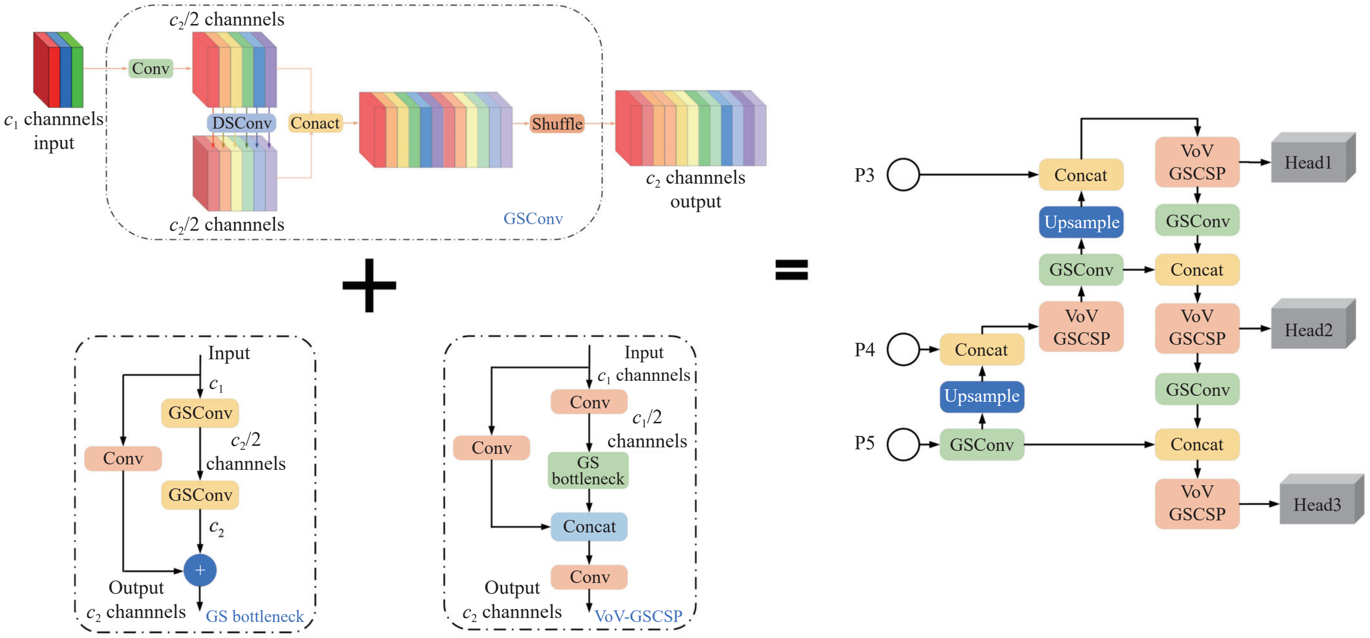


Figure 7 Structure of lightweight neck network

2.4.3 Loss function and activation function

In the actual picking environment, there is an overlap of lotus seedpods. If the bounding boxes with IoU values greater than the threshold are arbitrarily deleted, it may discard bounding boxes belonging to different targets, directly leading to missed detection. The CIoU_Loss (L_{CIoU}) is currently most commonly used as the loss function in YOLO algorithm[30], which is defined as follows:

$$L_{\mathrm{CIoU}} = 1 - \mathrm{IoU} + \frac{\rho^2(A,B)}{c^2} + \alpha v \tag{4}$$

$$\alpha = \frac{v}{(1-\mathrm{IoU})+v} \tag{5}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \tag{6}$$

where, IoU is the ratio of intersection and concatenation between the prediction frame and the real frame; $\rho(A, B)$ is the Euclidean distance between the center coordinates of the prediction frame and the real frame; c is the length of the diagonal of the external rectangle formed by the prediction frame and the real frame; α is the weight factor; v is the difference in aspect ratio between the prediction frame and the real frame; w, h, w^{gt}, and h^{gt} are the width and height of the prediction frame and the real frame, respectively.

However, the CIoU loss function relies too much on the aggregation of the bounding box regression metrics and does not take into account the direction of mismatch between the real frame and the predicted detection frame, which may lead to slower and less efficient convergence. In order to solve this problem, a more balanced loss function WIoU is introduced in this study, which uses the dynamic non-monotonic focusing mechanism of outlier instead of IoU to evaluate the quality of anchor frames, reducing the competitiveness of high-quality anchor frames while reducing the harmful gradients generated by low-quality examples, so that WIoU can focus on the common quality anchor frames and improve the whole performance of the detector. It is defined as follows:

$$L_{\mathrm{WIoU}} = rR_{\mathrm{WIoU}}L_{\mathrm{IoU}} \tag{7}$$

$$R_{\mathrm{WIoU}} = \exp\left(\frac{(x-x_{gt})^2+(y-y_{gt})^2}{(w_g^2+H_g^2)^*}\right) \tag{8}$$

where, (x_{gt}, y_{gt}) is the center coordinate of the enclosing frame; w_g and H_g are the width and height of the minimum enclosing frame, respectively; r is the non-monotonic focusing factor; R_{WIoU} is the normalized length of the centroid connection.

The use of activation functions allows the addition of nonlinear factors to the network and advances the model expression. The most commonly used activation functions in the YOLO family are Swish and Mish, which are defined as follows:

$$\mathrm{Swish} = x\mathrm{sigmoid}(\beta x) \tag{9}$$

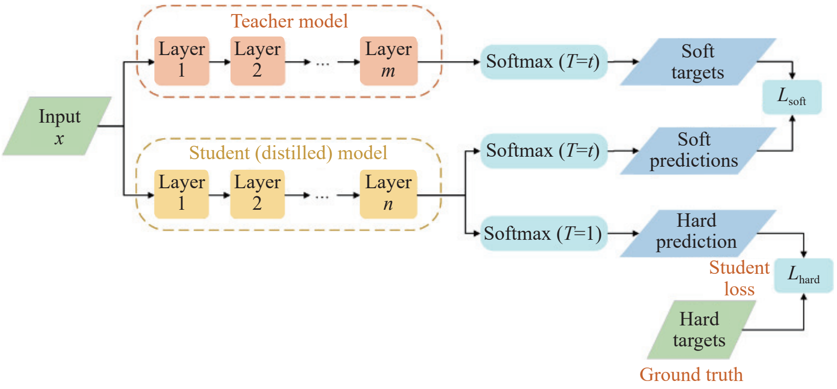
$$\mathrm{Mish} = x\tanh(\log(1+e^x)) \tag{10}$$

where, x is the parameter value passed in via the normalization layer; β is the variable coefficient.

Since the Swish activation function has problems such as large computational effort and unstable network performance, the Mish activation function was used to replace the Swish activation function in the backbone network. Its smooth characteristics can make the information penetrate deeply into the neural network, which makes the lotus seedpod detection more stable and accurate.

2.5 Knowledge distillation training strategy

In the lightweight process of the model, the performance of the detection model inevitably decreases as the number of parameters decreases. To compensate for the performance loss caused by the lightweight detection model, knowledge distillation is used to retrain the lightweight model, as shown in Figure 8.



Note: T and t are the distillation temperatures; L_{soft} is the loss function between the student model predictions and the teacher model predictions; L_{hard} is the loss function of the student model predictions to the true labels of the samples.

Figure 8 Structure of knowledge distillation

The enhanced lotus seedpods dataset is used to train a teacher network with deeper layers and stronger extraction ability, and then the probability prediction value of the lotus seedpods output by the teacher network is distilled at temperature T, and the predicted probability distribution of the lotus seedpods is obtained through the softmax layer as soft targets. At the same time, the probability prediction value of the lotus seedpods output by the student network is distilled at the same temperature T, and the predicted probability of the lotus seedpods is obtained after passing through the softmax layer, which is used as soft predictions. The loss function L_{soft} is defined as the follows:

$$L_{\mathrm{soft}} = -\sum_{j}^{N} p_j^T \log(q_j^T) \tag{11}$$

where, p_j^T and q_j^T are the jth class predicted probability output through the softmax layer at temperature T in the teacher network and student network, respectively; N is the total number of labels.

The teacher network may have a certain error rate. The possibility of errors propagated to the student network can be effectively reduced by using real labels as hard targets. L_{hard} is obtained by using the cross-entropy of the softmax output and hard targets of the student network under the condition of T=1, which is defined as the follows:

$$L_{\mathrm{hard}} = -\sum_{j}^{N} c_j \log(q_j^T) \tag{12}$$

where, c_j is the real label value of jth class.

The loss functions L_{hard} and L_{soft} are weighted together as the loss function L of the final distillation model so that the student model learns the teacher model while it is learning by comparison with the real label, which can effectively prevent the error information in the teacher network from being distilled into the student network. In this study, the YOLOv5m model is used as the teacher model, and the YOLOv5s with the above-improved method is used as the student model for distillation, thus improving the performance of the lightweight lotus seedpod detection model.

3 Results and analysis

3.1 Evaluation index

To validate the performance of the improved algorithm model, evaluation indexes such as Precision (P), Recall (R), F1-score, mAP, and FPS were used to evaluate the training model on the test dataset[31]. Intersection over Union (IoU)≥0.5 indicates a true case; IoU<0.5 indicates a false positive case; IoU=0 indicates a false negative case. IOU, precision, recall, mAP, and F1-score are defined as follows:

$$\mathrm{IoU}=\frac{|A\cap B|}{|A\cup B|} \tag{13}$$

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{14}$$

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{15}$$

$$\mathrm{mAP}=\frac{1}{k}\sum_{i=1}^{k}\mathrm{AP}_{\mathrm{i}} \tag{16}$$

$$\mathrm{F}_1\text{-score}=\frac{2P\cdot R}{P+R} \tag{17}$$

where, A is the prediction bracket; B is the true bracket; TP is the number of true positive cases; FP is the number of false positive cases; FN is the number of false negative cases; k is the number of detection categories; AP is the average precision.

3.2 Experimental platform and parameter setting

All experiments in this study are based on the Pytorch deep learning framework, programmed in Python, and run on Windows 11 with an Intel® Core™ i5-12400F (2.50 GHz) hexa-core CPU and NVIDIA GeForce RTX 3060 GPU. The image input size is 640×640 pixels and the initial learning rate of the model is 0.01. To speed up the training process and prevent overfitting, the momentum parameter is chosen to be 0.937. The model optimizer is SGD, the total number of epochs trained is 200, the batch size is 16, and the number of workers is 2. In order to reduce the training time, the migration learning method was used and the official pre-training weights were loaded, then the training of the model was started. The hyperparameters of the model training are listed in Table 2.

Table 2 Hyperparameters for model training

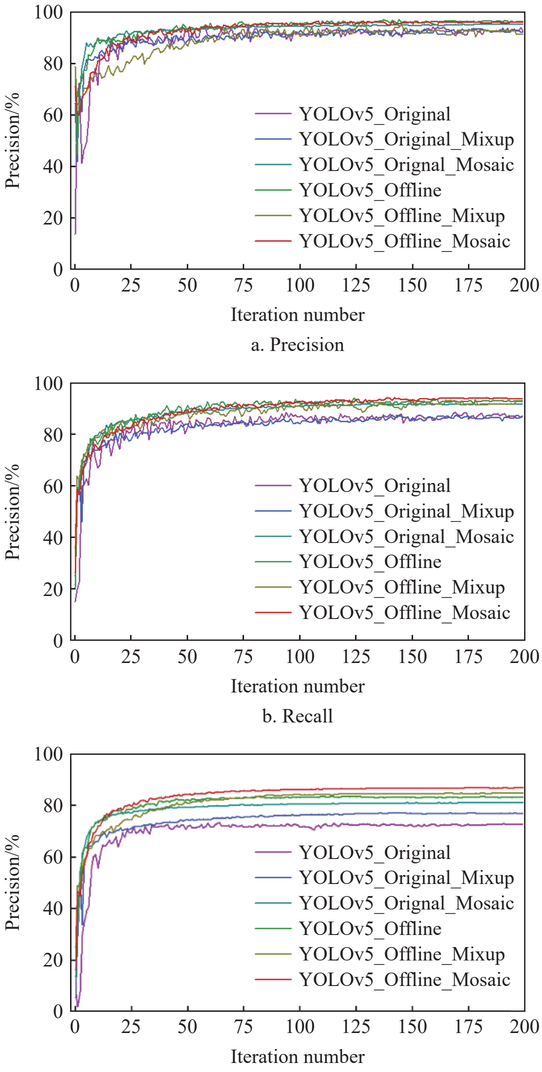
Parameters	Value
Input size/pixels	640×640
Learning rate	0.01
Momentum	0.937
Iterations	200
Batch size	16
Workers	2

3.3 Evaluation of the dataset augmentation

Using YOLOv5s as the basic framework of the lotus seedpod detection model, the effect on the validation set after each round of training with different amplification methods was analyzed, as shown in Figure 9. From the results, it can be seen that either Mix up or Mosaic online amplification with only the original dataset improves the mean average precision (mAP), and the results are mostly close in terms of precision and recall. Without any online augmentation, the offline augmentation-only approach also outperforms the model trained on the original dataset in terms of performance evaluation index, indicating that data augmentation can bring some improvement to the model performance.

After training the optimal model with different datasets, the model performance was tested on the test set and the results are listed in Table 3.

From Table 3, it can be seen that when using Mix up and Mosaic online augmentation strategies on the basis of the original dataset, the index mAP performance improves by 5.1 percent and 9.5 percent, respectively, while using the offline augmentation method, its mAP performance improves by 12.5%. When using Mix up and Mosaic online augmentation strategies on the basis of the offline augmentation dataset, the index mAP performance improves by 13.4 percent and 15.5 percent, respectively. Therefore, this study uses a combination of offline and online augmentation to maximize the data capacity and enrich the diversity of the dataset, thus achieving the greatest improvement in the mean average precision, as shown in Figure 10.



a. Precision

b. Recall

c. mean Average Precision (mAP)

Figure 9 Performance evaluation results of validation sets during training of different datasets

Table 3 Performance evaluation results of model test sets obtained from different training data

Datasets	Precision/%	Recall/%	mAP/%
Original	94.5	88.1	73.4
Original+Mix up	94.1	87.8	77.3
Original+Mosaic	95.4	91.7	81.1
Offline	97.1	93.8	83.9
Offline+Mix up	96.7	93.4	84.8
Offline+Mosaic	96.9	94.3	86.9

3.4 Experimental comparison before and after model improvement

In order to verify the effectiveness of the lightweight network model proposed in this study, the detection effects of different lightweight networks were compared, as listed in Table 4. It can be observed that GhostNet is larger than MobileNetv3 in terms of

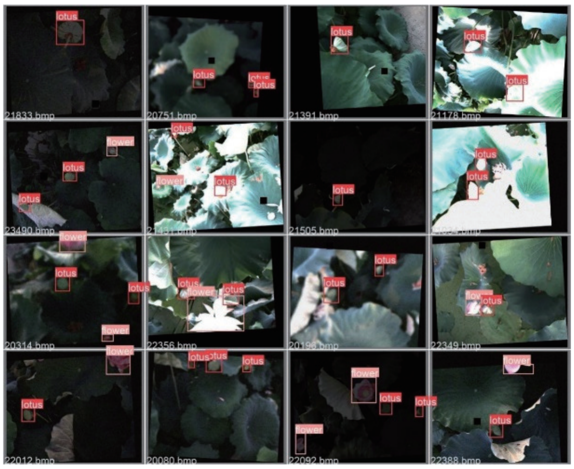


Figure 10 Offline and Mosaic combination of data enhancement

FLOPS with params, but it is superior to ShuffleNetv2 and MobileNetv3 in terms of mAP and detection frame rate (speed). Therefore, it verifies the effectiveness of using Ghost convolution instead of the original YOLOv5s model convolution. On this basis, the proposed lightweight bidirectional feature pyramid network is introduced. It can be found that after the introduction of BiFPN network structure, the mAP improved by 4.1%, but the params and FLOPs increased by 40.2% and 32.8% compared with the previous ones. However, with the introduction of the GSConv and VoV-GSCSP modules, the params and FLOPS decreased by 24.4% and 28.6%, respectively. Therefore, the lightweight network model proposed in this study is effective and can reduce the params and FLOPs of the model while ensuring detection accuracy.

Table 4 Comparative experiments of different lightweight networks

Model	mAP/ %	Params/ $\times10^6$	FLOPs/ $\times10^9$	F1-score/ %	Size/ MB	Speed/ fps
YOLOv5s_ShuffleNetv2	80.6	3.80	8.2	90.9	7.8	9.1
YOLOv5s_MobileNetv3	79.8	3.55	6.3	91.8	7.3	9.2
YOLOv5s_GhostNet	81.2	3.68	8	92.3	7.6	8.8
YOLOv5s_ G-BiFPN	85.3	6.15	11.9	93.4	12.4	8.6
YOLOv5s_GBGV	84.6	4.65	8.5	92.4	9.5	9.6

Based on the improved network, the LeakyReLu, Mish, HSwish activation functions, and CIoU and WIoU loss functions were compared in the experiments. As listed in Table 5, Mish is not as fast as LeakyReLu and HSwish in terms of speed, but its mAP is better; while CIoU and WIoU are similar in terms of speed, but the mAP of WIoU is better than CIoU. So the combination of Mish activation function and WIoU loss function was chosen.

Table 5 Comparison results of different activation functions and IoU loss under the same model

Activation/IoU Loss	mAP/%	Params/$\times10^6$	F1-score/%	Speed/fps
LeakyReLu/CIoU	83.2	4.65	92.0	9.2
HSwish/CIoU	83.6	4.65	92.2	9.1
Mish/CIoU	84.2	4.65	92.3	10.6
LeakyReLu/WIoU	84.1	4.65	91.5	9.4
HSwish/WIoU	84.4	4.65	93.0	9.3
Mish/WIoU	85.5	4.65	93.4	10.9

In order to verify whether all the improved methods based on the YOLOv5s algorithm model can improve the detection performance, we conducted an ablation study of all the improvements, and the experimental results are listed in Table 6.

Table 6 Comparison of ablation study

Model	mAP/ %	Params/ $\times10^6$	FLOPs/ $\times10^9$	F1-score/ %	Size/ MB	Speed/ fps
YOLOv5s	86.9	7.01	15.8	95.5	14.1	10.5
YOLOv5s_GBGV	84.6	4.65	8.5	92.4	9.5	9.6
YOLOv5s_GBGV_M/W	85.5	4.65	8.5	93.4	9.5	10.9
YOLOv5s_GBGV_M/W+KD	89.7	4.65	8.5	96.2	9.5	14.4

Compared with the original YOLOv5s model, when using the proposed lightweight detection network algorithm model, the params and FLOPS decrease by 33.7% and 46.2%, respectively. After using the Mish activation function and WIoU loss function, the mAP improves by 1.1 percent, and the other parameters remain unchanged. With the knowledge distillation training strategy, the mAP of the model improves by 4.7%. From the results of the ablation experiments, all the improvements made in this study based on YOLOv5s have played their proper roles compared with the original model.

To visualize the effectiveness of the proposed improved method for the lightweight lotus seedpod detection model, the heat maps were used to visualize the extraction of feature maps and target localization, as shown in Figure 11. Compared with the YOLOv5s model, the proposed lightweight lotus seedpod detection algorithm model can successfully extract the features of the lotus seedpods and make correct predictions. The detection results of the model after distillation training are more based on the lotus seedpods themselves and less dependent on the external environment. Therefore, the model is less affected by the external environment, making it more robust and semantically informative.

3.5 Experiment comparison with different detection algorithm models

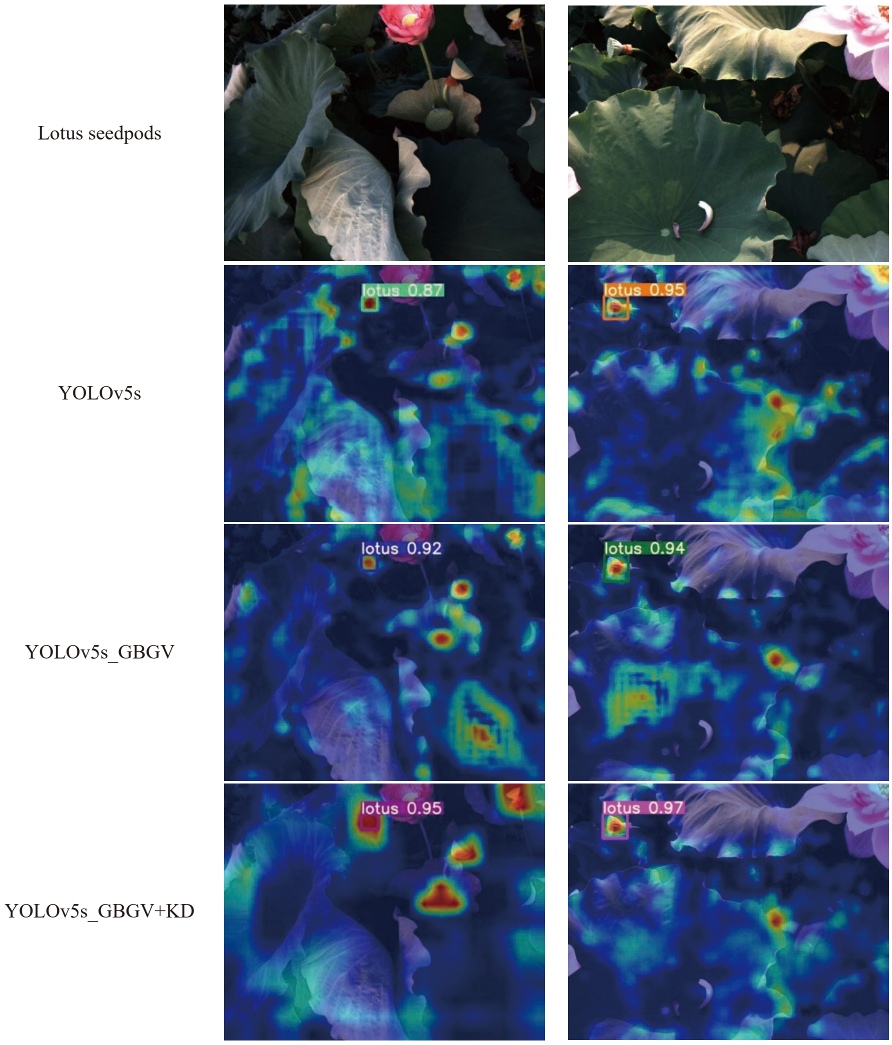
The proposed lightweight lotus seedpod detection algorithm model is compared with the YOLOv5s detection algorithm model, which reduces the params and FLOPs of the model while improving the mAP. The effect of lotus seedpods detection before and after the model improvement is shown in Figure 12.

The comparison with the state-of-the-art algorithm models at the present stage was carried out, and the comparison results are shown in Table 7. It can be found that compared with Faster-RCNN, SSD, and YOLOv7-tiny algorithm models, the present algorithm model has lower values of the params and FLOPs, which effectively reduces the computational effort of the model. In the process of lotus seedpod detection, the precision, recall, and mAP values of this algorithm model are better than those of other models, and the size of this model is smaller than that of other detection models. It further verifies the effectiveness of the improved method proposed in this study.

4 Conclusions

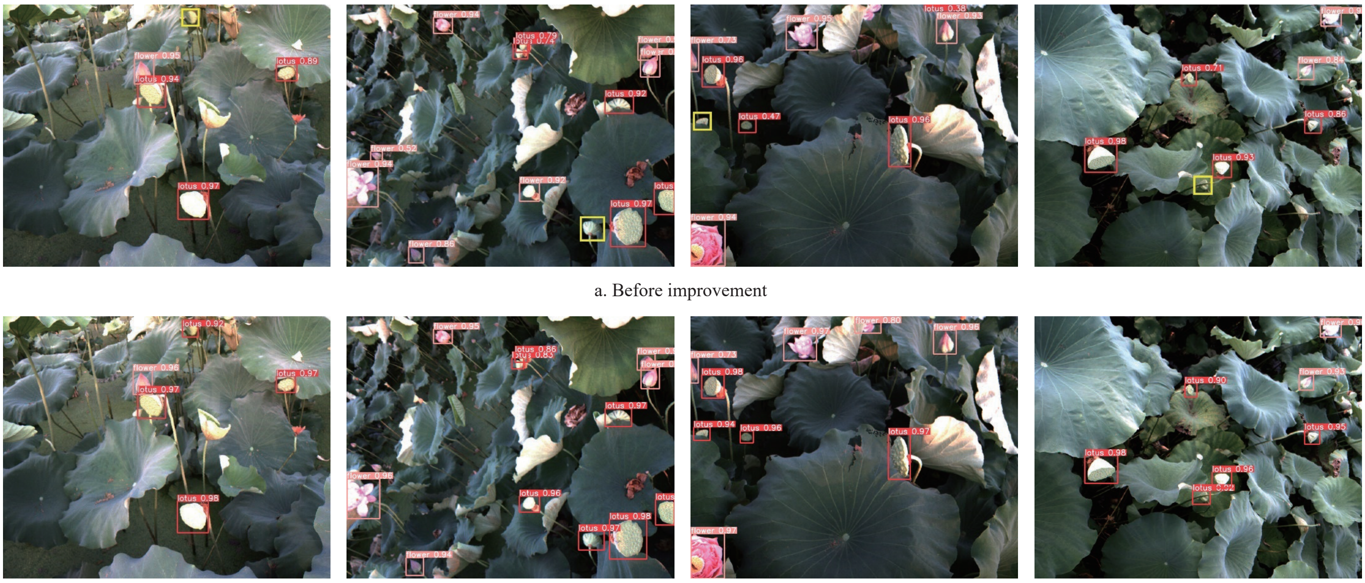
In order to apply the detection recognition model to the actual lotus seedpod harvesting environment, a lightweight lotus seedpod detection method based on the YOLOv5s model was proposed. The effectiveness of the proposed method was verified through ablation study. The following conclusions can be drawn:

1) The combination of offline and online augmentation (Mosaic) was used to augment the lotus seedpods dataset, which greatly improved the training accuracy and increased mAP by 15.5%. GhostNet was used as the backbone network, and a lightweight bidirectional feature pyramid network was proposed, which significantly reduced the params and FLOPs of the model by 33.7% and 46.2%, respectively. The combination of WIoU loss function and Mish activation function was adopted, which enabled



a. Detection example 1 b. Detection example 2

Figure 11 Visualization feature map of the improved model



a. Before improvement

b. After improvement

Note: indicates missing detection.

Figure 12 Comparison of actual detection effect before and after model improvement

the model to increase the training convergence speed as well as improve the detection precision. The knowledge distillation training strategy was used to make the lightweight lotus seedpod detection model have the learning capability of a complex network model;

2) The proposed model was evaluated through ablation study, including various performance evaluation metrics and visual feature

Table 7 Comparison of the results of the most advanced detectors at this stage

Model	Precision/ %	Recall/ %	mAP/ %	Params/ $\times10^6$	FLOPs/ $\times10^9$	Size/ MB	Speed/ fps
Faster-RCNN	68.2	59.9	60.0	137.1	370.2	110.8	11.3
SSD	94.9	48.6	60.4	26.3	62.8	93.3	78.3
YOLOv7-tiny	94.5	87.8	79.8	6.02	13.2	11.98	9.5
YOLOv5s	96.7	94.3	86.9	7.01	15.8	14.1	10.5
Method proposed in this study	97.5	94.9	89.7	4.65	8.5	9.5	14.4

maps. The experimental results show that the proposed method achieves the expected effect, with a precision of 97.5%, recall of 94.9%, mAP of 89.7%, params of 4.65M, FLOPs of 8.5G, and model size of 9.5M. Compared with other mainstream detection models, the proposed model has the advantages of less computation, smaller model size, and higher detection accuracy, which is beneficial to deploying the detection model on mobile terminals with limited computational power and small storage space, and provides theoretical research and technical support for intelligent picking operations of lotus seedpods;

Since the present model does not classify the lotus seedpods for maturity detection, future work will study the lotus seedpods classification harvesting model to further improve the detection speed and accuracy of lotus seedpods.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 32171899), Zhejiang Provincial Natural Science Foundation of China (Grant No. LGN22C130006), and the development of technology for lotus seedpods identification and localization and maturity detection (Grant No. 22020223-J). Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of Zhejiang Sci-Tech University.

[References]

[1] Zhu F L, Sun H, Wang J, Zheng X W, Wang T, Diao Y, et al. Differential expression involved in starch synthesis pathway genes reveal various starch characteristics of seed and rhizome in lotus (*Nelumbo nucifera*). Journal of Food Science, 2022; 87(9): 4250–4263.

[2] Bangar S P, Dunno K, Kumar M, Mostafa H, Maqsood S. A comprehensive review on lotus seeds (*Nelumbo nucifera* Gaertn.): Nutritional composition, health-related bioactive properties, and industrial applications. Journal of Functional Foods, 2022; 89: 104937.

[3] Chen C J, Li G T, Zhu F. A novel starch from lotus (*Nelumbo nucifera*) seeds: Composition, structure, properties and modifications. Food Hydrocolloid, 2021; 120: 106899.

[4] Zhang X, He L, Majeed Y, Whiting M D, Karkee M, Zhang Q. A precision pruning strategy for improving efficiency of vibratory mechanical harvesting of apples. Transactions of the ASABE, 2018; 61(5): 1565–1576.

[5] Lytridis C, Kaburlasos V G, Pachidis T, Manios M, Vrochidou E, Kalampokas T, et al. An overview of cooperative robotics in agriculture. Agronomy, 2021; 11(9): 1818.

[6] Rose D C, Lyon J, de Boon A, Hanheide M, Pearson S. Responsible development of autonomous robotics in agriculture. Nature Food, 2021; 2(5): 306–309.

[7] Li H P, Li C Y, Li G B, Chen L X. A real-time table grape detection method based on improved YOLOv4-tiny network in complex background. Biosystems Engineering, 2021; 212: 347–359.

[8] Zhang X H, Toudeshki A, Ehsani R, Li H L, Zhang W F, Ma R J. Yield estimation of citrus fruit using rapid image processing in natural background. Smart Agricultural Technology, 2022; 2: 100027.

[9] Koirala A, Walsh K B, Wang Z, Mccarthy C. Deep learning for real-time fruit detection and orchard fruit load estimation: Benchmarking of '*Mangoyolo*'. Precision Agriculture, 2019; 20(6): 1107–1135.

[10] Lu Y Z, Young S. A survey of public datasets for computer vision tasks in precision agriculture. Computers and Electronics in Agriculture, 2020; 178: 105760.

[11] Sultana F, Sufian A, Dutta P. A review of object detection models based on convolutional neural network. Intelligent Computing: Image Processing Based Applications, 2019; pp.1-16.

[12] Zhou Z, Majeed Y, Naranjo G D, Gambacorta E M T. Assessment for crop water stress with infrared thermal imagery in precision agriculture: a review and future prospects for deep learning applications. Computers and Electronics in Agriculture, 2021; 182: 106019.

[13] Pathak H, Igathinathane C, Howatt K, Zhang Z. Machine learning and handcrafted image processing methods for classifying common weeds in corn field. Smart Agricultural Technology, 2023; 5: 100249.

[14] Zhang Z, Igathinathane C, Flores P, Ampatzidis Y, Liu H, Mathew J, et al. Time effect after initial wheat lodging on plot lodging ratio detection using UAV imagery and deep learning. Unmanned Aerial Systems in Precision Agriculture, 2022; pp.50-72.

[15] Jiao L C, Zhang F, Liu F, Yang S Y, Li L L, Feng Z X, et al. A survey of deep learning-based object detection. IEEE Access, 2019; 7: 128837–128868.

[16] Din A, Ismail M Y, Shah B B, Babar M, Ali F, Baig S U. A deep reinforcement learning-based multi-agent area coverage control for smart agriculture. Computers and Electrical Engineering, 2022; 101: 108089.

[17] Zhang Z, Flores P, Friskop A, Liu, Z H, Igathinathane, C, Han, X, et al. Enhancing wheat disease diagnosis in a greenhouse using image deep features and parallel feature fusion. Front Plant Sci, 2022; 13: 834447.

[18] Lu Y Z, Lu R F, Zhang Z. Detection of subsurface bruising in fresh pickling cucumbers using structured-illumination reflectance imaging. Postharvest Biol Tec, 2021; 180: 111624.

[19] Flores P, Zhang Z, Igathinathane C, Jithin M, Naik D, Stenger J, et al. Distinguishing seedling volunteer corn from soybean through greenhouse color, color-infrared, and fused images using machine and deep learning - ScienceDirect. Ind Crop Prod, 2020; 161: 113223.

[20] Wang P, Niu T, Mao Y R, Zhang, Z, Liu, B, He, D J. Identification of apple leaf diseases by improved deep convolutional neural networks with an attention mechanism. Frontiers Plant Science, 2021; 12: 723294.

[21] Kang H W, Chen C. Fruit detection, segmentation and 3d visualisation of environments in apple orchards. Computers and Electronics in Agrulture, 2020; 171: 105302.

[22] 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. Computers and Electronics in Agriculture, 2019; 157: 417–426.

[23] Ma J, Lu A E, Chen C, Ma X D, Ma Q C. YOLOv5-lotus an efficient object detection method for lotus seedpod in a natural environment. Computers and Electronics in Agriculture, 2023; 206: 107635.

[24] Bhagat S, Kokare M, Haswani V, Hambarde P, Kamble R. WheatNet-lite: A novel light weight network for wheat head detection. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal: IEEE, 2021; pp.1332–1341.

[25] Zha M F, Qian W B, Yi W L, Hua J. A lightweight yolov4-based forestry pest detection method using coordinate attention and feature fusion. Entropy, 2021; 23(12): 1587.

[26] Cui M D, Lou Y Y, Ge Y L, Wang K Q. LES-YOLO: A lightweight pinecone detection algorithm based on improved YOLOv4-Tiny network. Computers and Electronics in Agriculture, 2023; 205: 107613.

[27] Zhang Y, He S P, Wa S Y, Zong Z Q, Liu Y L. Using generative module and pruning inference for the fast and accurate detection of apple flower in natural environments. Information, 2021; 12(12): 495.

[28] Jiang P Y, Ergu D, Liu F Y, Cai Y, Ma B. A review of yolo algorithm developments. Procedia Computer Science, 2022; 199: 1066–1073.

[29] Hu W X, Xiong J T, Liang J H, Xie Z M, Liu Z Y, Huang Q Y, et al. A method of citrus epidermis defects detection based on an improved YOLOv5. Biosystems Engineering, 2023; 227: 19–35.

[30] Ji S J, Ling Q H, Han F. An improved algorithm for small object detection based on YOLO v4 and multi-scale contextual information. Computers and Electrical Engineering, 2023; 105: 108490.

[31] Lu Z H, Zhao M F, Luo J, Wang G H, Wang D C. Design of a winter-jujube grading robot based on machine vision. Computers and Electronics in Agriculture, 2021; 186: 106170.