IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2022.Doi Number

Research on the detection method of safflower filaments in natural environment based on improved YOLOv5s

Bangbang Chen[1,2], Feng Ding[1,*], Baojian Ma[2], Xiangdong Liu[2] and Shanping Ning[1]

[1]School of Mechatronic Engineering, Xi'an Technological University, Xi'an 710021, China
[2]School of Mechatronic Engineering, Xinjiang Institute of Technology, Aksu843100, China

Corresponding author: Feng Ding (e-mail: dfeng@xatu.edu.cn).

This work was supported by the second batch of Tianshan Talent Cultivation Plan for Young Talent Support Project 2023TSYCQNTJ0040, and the Natural Science Basic Research Program of Shaanxi 2023-JC-YB-347.

ABSTRACT The accurate and rapid identification of safflower filaments is a prerequisite for automating harvesting. This paper proposes a lightweight, high-precision detection model for safflower filaments based on YOLOv5s, named YOLOv5s-MCD, to address the issues of large existing network model sizes and low detection accuracy in complex natural environments. The Backbone of the YOLOv5s-MCD model was optimized into a lightweight improved network MobileNetv2 with DSC and CA modules, and the neck part incorporated the CA attention mechanism. The loss function is improved from DIoU's non-maximum suppression method to CIoU to reduce the model size and improve the detection accuracy and speed. The experimental results show that the size of the YOLOv5s-MCD model is its size by 7.69 MB compared to the original YOLOv5s model, with a mean Average Precision (mAP) of 95.6% and an average detection time of only 3.2ms per image. When tested under unobstructed, obstructed, backlighting, shaking, and wide-angle natural environments, the improved YOLOv5s-MCD model increased the mAP value by 4.4, 0.7, 3.3, 3.4, and 1.0 percentage, respectively, compared with the YOLOv5s model, with improved F1 scores and confidence levels. This indicates that the improved lightweight model can achieve fast, real-time, and accurate detection of the safflower filaments in complex environments. The research results can provide a technical reference for the development of field safflower filament-harvesting robots.

INDEX TERMS Safflower filament, Natural environment, Target detection, YOLOv5s, Lightweight.

I. INTRODUCTION

Safflower (Carthamus tinctorius L.), also known as red bluebottle or thistle safflower, promotes blood circulation to remove blood stasis, regulates menstruation, relieves pain, and lowers blood lipid and blood pressure [1]. With the continuous increase in market demand for safflower, the safflower market is rapidly developing towards industrialization and commercialization, playing an immeasurable role in regional economic development. However, in the entire safflower industry chain, the harvesting of safflower filaments has not yet been mechanized and mainly relies on manual labor, leading to significant waste due to untimely harvesting, which severely restricts the large-scale development of the industry [2-4]. With the development of artificial intelligence and deep learning technology, automation of safflower filament harvesting has become possible. However, safflower unstructured field planting presents a variety of growth postures and different flowering periods, and the characteristics of the filaments, such as shape, size, and color, also change with time and light conditions. These factors inevitably lead to mutual shading among safflowers and overlapping of targets, posing a pressing issue for the development of intelligent safflower filament picking robots: how to quickly and accurately identify small safflower filament targets in such complex natural environments.

Currently, the recognition of safflower filaments has attracted the attention of researchers. For instance, Zhang Zhenguo et al. [5-6] conducted recognition studies on safflower filaments in natural environments based on improved YOLOv3 and Faster R-CNN, showing that the average precision mean of YOLOv3 reached 91.89%, with an average detection speed of 51.1 frames /s, and the average precision mean of Faster R-CNN reached 91.49%. Wang Xiaorong et al. [7] identified safflower filaments in

complex environments based on YOLOv7 by adding a Swin Transformer attention mechanism and improving the Focal Loss function; the test results indicated an average detection precision of 88.5%, without mentioning the detection speed per single image. Scholars have also studied the identification of safflower filaments through image segmentation methods; for example, Dong Funan et al. [8] used the Otsu method to establish a target segmentation method for safflower filaments, completing the target segmentation quickly and professionally through color space conversion, bilateral filtering, and the least squares principle, achieving an accuracy rate of 90%.

Although rapid and accurate recognition technologies for safflower filaments are still at an early stage of research, the identification of various flowers and small fruit targets has already been widely applied. For example, Fan Xiangpeng et al. [9] proposed an improved Faster R-CNN and data augmentation recognition method to address the low recognition rate and poor robustness of existing models for weed accompanying cotton seedlings, achieving an average recognition time of 0.261s for single images in natural environments and an average precision of 94.21%, which shows a clear advantage over other algorithms. Shang Yuying et al. [10] detected apple blossoms in natural scenes based on YOLOv5s, with test results showing an average precision of 97.2%, a detection speed of 60.17 frames/s, and a model size of 14.09 MB, demonstrating high robustness. Similarly, Wu Dihua et al. [11], based on an improved channel-pruned YOLOv4 model for real-time recognition of apple blossoms, showed that the average precision reached 97.31%, with a detection speed of 72.33 frames /s, and a model size of 12.46 MB. Chen Chunlin et al. [12] designed a computer vision system for identifying tea leaf picking points based on YOLOv3, using the Mobilenetv2 algorithm for classification and combining deep learning with traditional image processing algorithms, achieving a picking point location accuracy of 83%; Li Jie [13] also addressed the problem of difficult precise identification in complex tea leaf environments, proposing a high-precision lightweight detection model based on an improved YOLOv4, replacing the backbone of YOLOv4 with GhostNet and introducing the CBAM attention mechanism to enhance the model's feature extraction capability, with test results showing an accuracy rate of 85.15% and a parameter reduction of 82.36%, providing a foundation for precise tea leaf picking. Miao Ronghui et al. [14] improved the detection speed by changing the backbone network of YOLOv7 without reducing the accuracy of cherry tomato detection, and the results indicated an average detection time of 82ms per single image, a model size of 66.5 MB, and an average precision of 98.2%. In summary, although deep learning technology has been successfully applied in fields such as cotton, tea leaves, and apple blossoms [15-17], there are few deployments and applications of recognition systems for safflower filament harvesting. The main reasons are the strong dependence of deep learning models on high-performance computing platforms and the impact of non-structured planting patterns of safflower filaments in natural environments, including filament shading, multi-target overlap, and changes in illumination, which result in existing target detection models having low accuracy, large parameter computation, and high memory occupation of model weights. Therefore, it is necessary to study strategies to significantly reduce the model size while ensuring unchanged or slightly changed detection accuracy, preparing for the deployment of the model on low-computing power mobile platforms in the field.

Based on this, this study focuses on the issues of low target detection accuracy and large model size of existing models for safflower filaments in natural scenes, and conducts research on a lightweight model for safflower filament target detection based on the YOLOv5s network architecture. By introducing the lightweight network structure Mobilenetv2 into the Backbone network for improvement, this study achieves model lightweighting by adding the lightweight attention mechanism CA in the Neck network and using DIoU's non-maximum suppression method in the CIoU loss function, which improves the model's detection accuracy and convergence speed. In addition, this study also compares different backbone network structures, attention mechanisms, and loss functions of YOLOv5s variant models to evaluate the performance of the improved model. The research results can provide important technical support for the development of automatic safflower filament harvesting robots.

II. Materials and Methods

A. COLLECTION OF SAFFLOWER FILAMENT IMAGE DATA

This paper focuses on the precious medicinal herb "Yumin Thornless Safflower" which is extensively cultivated in Xinjiang, China. The safflower filament images used were obtained from the Labor Education Practice Base of Xinjiang University of Technology in the Aksu region, Xinjiang. The image capturing device was a Huawei Nova 7 Pro with a resolution of 4608×3456 pixels. Based on the flowering growth characteristics of the safflower filaments, the image collection period was from July 10 to August 1, 2023. The camera angles combined 90-degree horizontal shots and 135-degree overhead shots. Data were collected under different natural conditions such as sunny, cloudy, and overcast weather, during the time periods of 9:00-10:00 AM, 3:00-4:00 PM, and 9:30-10:30 PM. The dataset included safflower filament images with different lighting, varying obstructions of leaves and filaments, varying obstructions of fruits and filaments, and varying obstructions of filaments and filaments. Figure 1 shows examples of safflower filament

images collected under different natural conditions. A total of 3000 safflower filament images were collected and stored in the JPEG format.

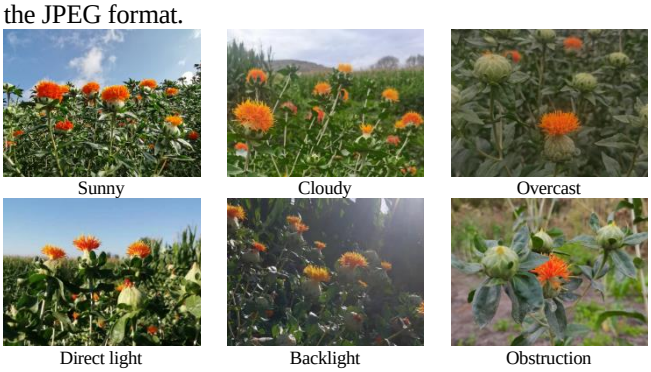


FIGURE 1.Images of safflower filaments in different natural environments.

B. SAFFLOWER FILAMENTS IMAGE DATASET CONST-RUCTION AND ANALYSIS

To enhance the robustness and generalizability of the safflower filament recognition model for more accurate feature extraction under various environmental conditions, this study applied two data augmentation strategies brightness adjustment and blur processing to the original image dataset after comprehensive analysis of field operation environments. Brightness adjustment simulates the effects of changing light conditions due to different weather conditions, whereas blur processing aims to mimic the blurring effect caused by the movement of harvesting machinery, as exemplified in Figure 2. Using these methods, the number of image samples was increased to 4500, which were then divided into training (3150 images), validation (900 images), and test sets (450 images) at a ratio of 7:2:1 for model training and evaluation. Furthermore, the dataset adhered to the PASCAL VOC formatting standards, and the safflower filaments within the images were precisely annotated using LabelImg software. During the annotation process, operations were performed only on the minimum bounding rectangles of each filament, excluding instances that were too obscured, blurry, or distant to be recognized. After annotation, the information was saved in an XML file. Considering the specific conditions encountered by safflower harvesting robots during field operations, this study categorized the safflower filament targets into two classes: "Safflower-B" and "Safflower-D," based on which the data were classified and annotated.

FIGURE 2.Example of safflower filaments image enhancement.

C. IMPROVED YOLOV5S DETECTION MODEL

1) YOLOV5S NETWORK MODEL

YOLOv5 is a single-stage object detection algorithm that offers five different network model variants of varying sizes [18-19]: YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. These models are designed to optimize the balance between computational efficiency and detection performance and have introduced the focus module and Spatial Pyramid Pooling (SPP) [20] structure into the network architecture to further enhance feature extraction capabilities compared to their predecessor, YOLOv4. As the number and depth of the network layers increase, the number of parameters correspondingly increases. Considering the specific requirements for the field harvesting of safflower filaments, this study selected the YOLOv5s model. This model strikes an optimal balance between model size and detection speed while maintaining the necessary detection accuracy. The architecture of the YOLOv5s model primarily consists of Backbone and Neck components.

The Backbone part is mainly used for feature extraction of safflower filament images, using CSPDarknet53 as the backbone network. The input size of the safflower filament image is 640×640×3. Through the Focus layer, the input image is divided into four 320×320×3 slices, and then the Concat operation outputs a feature map of 320×320×12. After a convolution operation with 32 kernels, it becomes a feature map of 320×320×32, thereby achieving the effect of reducing computation and accelerating training speed; layers 1, 3, 5, and 7 are convolutional layers; layers 2, 4, 6, and 8 are CSP layers. This module refers to the skip connection idea of the ResNet model to obtain richer semantic information [21], the SPPF module is designed by referring to the spatial pyramid pooling module (SPP), realizing multi-scale information fusion to improve model performance while speeding up model processing. The Neck part uses the feature fusion method of Path Aggregation Network (PANet), including the top-down FPN module and the bottom-up PAN module, which better integrates features of different scales and enhances the detection accuracy of the model. It has 15 sub-layers, and the outputs from the 4th and 6th layers of the backbone module are concatenated and fused with the 15th and 11th layers of the Neck module to output richer feature information.

2) COORDINATE ATTENTION MODULE

To enhance the fusion processing performance of the Neck module in extracting features from different levels of the backbone network and to improve the recognition accuracy and speed of the entire network model, an attention mechanism was embedded in the network for optimization and improvement. The currently prevalent lightweight attention mechanisms include SE (Squeeze and Excitation) [22], CBAM (Convolutional Block Attention Module) [23], ECA (Efficient Channel Attention) [24], and CA (Coordinate Attention) [25]. The SE attention mechanism considers only inter-channel information and enhances the feature selection capability while neglecting the importance of spatial information. CBAM combines channel and spatial attention

mechanisms by reducing the input tensor and introducing large-scale convolutional kernels to extract spatial features, thus ignoring long-range dependencies and the increasing computational complexity. ECA, proposed by Wang et al. in 2020, primarily aims to simplify the channel attention mechanism and reduce parameter count and computational complexity, but does not integrate spatial information, limiting its effectiveness in capturing global context and channel-spatial relationships. The CA attention mechanism, introduced by Hou et al. in 2021, enhances feature representation by innovatively combining channel attention and spatial coordinate information, thereby improving the model's ability to capture both global and local features with high computational efficiency. It effectively retained the spatial location information of the safflower filament, significantly enhancing the recognition capability of the network model. Based on this analysis, this optimization employs a lightweight CA attention mechanism module, the structure of which is illustrated in Figure 3.

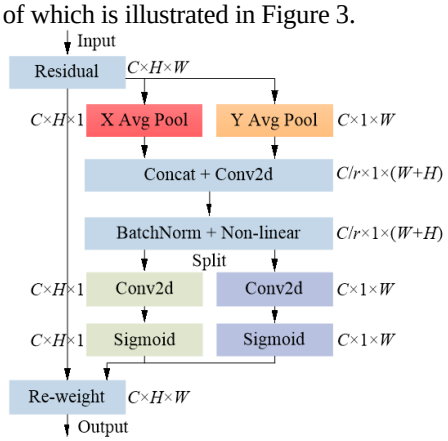


FIGURE 3.CA attention mechanism network structure.

As shown in Figure 3, the CA mechanism enhances the ability of mobile networks to learn feature representations by embedding coordinate information and generating coordinate attention in two steps: encoding precise positional information for channel relationships and long-range dependencies. Coordinate information embedding performs channel-wise average pooling along the X and Y directions on the input feature map of size $C\times H\times W$, generating feature maps of sizes $C\times H\times 1$ and $C\times 1\times W$ respectively. This helps the attention module capture long-range dependencies with precise positional information, allowing the network to locate objects of interest more accurately. The operations are expressed in equations (1) and (2):

$$z_c^h(h)=\frac{1}{W}\sum_{0\le i<W}x_c(h,i) \tag{1}$$

$$z_c^w(w)=\frac{1}{H}\sum_{0\le j<H}x_c(j,w) \tag{2}$$

Where $z_c^h(h)$ is the output of the c-th channel at height h, and $z_c^w(w)$ is the output of the c-th channel at width w.

To better capture the global receptive field and encode precise positional information, coordinate attention was generated using representations produced by embedding coordinate information. By spatially concatenating Equations (1) and (2), and then applying a 1×1 convolutional transformation function F_1 and a nonlinear activation function to obtain an intermediate feature map [25-26],

$$\mathrm{f}=\delta(F_1([z^h,z^w])) \tag{3}$$

where, f represents the intermediate feature map encoding spatial information in both the horizontal and vertical directions, and $\mathrm{f}\in\mathbf{R}^{C/r\times(H+W)}$, where r controls the reduction ratio of the block size. δ is a nonlinear activation function, and $[z^h,z^w]$ is the spatial dimension concatenation. Subsequently, along the spatial dimension, f is split into two separate tensors $\mathrm{f}^h\in\mathbf{R}^{C/r\times H}$ and $\mathrm{f}^w\in\mathbf{R}^{C/r\times W}$. Two 1×1 convolutional transformation functions F_h and F_w are used to transform f^h and f^w into tensors with the same channels as input feature X, resulting in

$$\mathrm{g}^h=\sigma(F_h(\mathrm{f}^h)) \tag{4}$$

$$\mathrm{g}^w=\sigma(F_w(\mathrm{f}^w)) \tag{5}$$

In this formula, σ is a sigmoid activation function. Finally, g^h and g^w are expanded and used as attention weights to obtain the output Y of the CA attention module.

$$y_c(i,j)=x_c(i,j)\times g_c^h(i)\times g_c^w(j) \tag{6}$$

3) IMPROVED MOBILENETV2 MODULE

To ensure a significant reduction in the model parameters and size while maintaining high accuracy in detecting safflower filaments, we conducted a comparative analysis of four lightweight networks: GhostNet[27], ShuffleNetV2[28], MobileNetV2[29], and MobileNetV3[30]. The GhostNet module, despite utilizing channel-wise grouped convolution under resource-constrained conditions, increases computational complexity. ShuffleNetV2's deeper network architecture increases computational costs without markedly improving small object feature detection. Both MobileNetV2 and MobileNetV3 leverage depthwise separable convolutions and residual modules. However, MobileNetV3, which incorporates the SE attention mechanism, fails to account for the spatial information of overlapping safflower filaments, and is larger MobileNetV2. Consequently, the MobileNetV2 module is adopted in this study. The improved MobileNetV2 network architecture is illustrated in Figure 4. The expansion factor of the input size is denoted by t, c represents the number of channels in the output feature map, and s stands for stride (applied only to the first layer; all others are set to 1). The improved MobileNetv2 network model incorporates the CA attention mechanism and continues the design philosophy of Depthwise Separable Convolution (DSC) [31], Linear Bottleneck, and Inverted Residual modules in the network structure.

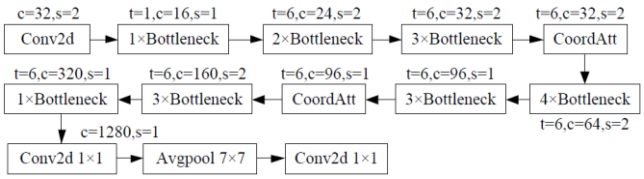


FIGURE 4.Improved MobileNetV2 Network Structure.

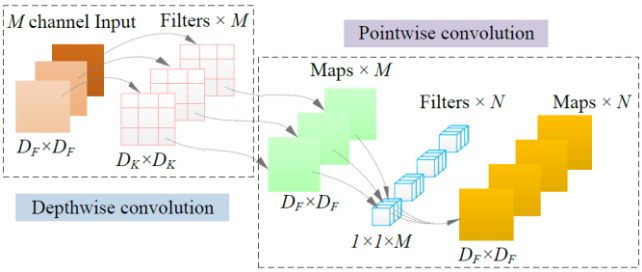


FIGURE 5.Depthwise Separable Convolution.

Depthwise separable convolution consists of depthwise and pointwise convolutions. It first uses depthwise convolution to gather information on the feature channels and then uses pointwise convolution to complete the output of the feature information, as shown in Figure 5. Unlike standard convolutions (SC), the kernels in depthwise convolution are single-channel, meaning a single kernel is applied to each channel to perform operations, resulting in the same number of input and output feature map channels, which greatly reduces the computational load and number of parameters. Pointwise convolution uses 1×1 kernels to perform dimension-raising operations, constructing new features through linear combinations of input channels with N 1×1 kernels, thus addressing the issue of having too few feature maps. In a depthwise separable convolution network, assuming that the size of the depthwise convolution kernel is $D_K \times D_K$, M is the number of input channels, N is the number of output channels, and $D_F \times D_F$ is the output feature map size, the computational cost of depthwise separable convolution and standard convolution can be represented by equations (7) and (8), respectively. The reduction in the computational cost for depthwise separable convolution is shown in equation (9). In MobileNetv2, $D_K = 3$; therefore, the computational cost is eight to nine times lower than that of standard convolution, with only a slight decrease in accuracy.

$$D_K \times D_K \times M \times D_F \times D_F + M \times N \times D_F \times D_F \quad (7)$$

$$D_K \times D_K \times M \times N \times D_F \times D_F \quad (8)$$

$$\frac{D_K \times D_K \times M \times D_F \times D_F + M \times N \times D_F \times D_F}{D_K \times D_K \times M \times N \times D_F \times D_F} = \frac{1}{N} + \frac{1}{D_K{}^2} \quad (9)$$

In the classic residual structure, a pattern of reducing the dimensions before increasing them is adopted. However, MobileNetV2 optimizes this into an inverted residual structure that first increases and then reduces the dimensions, as shown in Figure 6. The inverted residual structure not only ensures the extraction of high-dimensional feature information but also reduces the number of parameters and computational cost. To address the issue of feature loss caused by the use of the nonlinear activation function ReLU in the last 1×1 convolutional layer of the inverted residual structure, the ReLU activation function is improved to a linear activation function, which significantly reduces the loss of information when passing through narrow layers, hence it is called linear bottleneck convolution, as shown in Figure 7. It should be noted that shortcut connections are present only when *s*=1, and the input and output feature maps have the same shape.

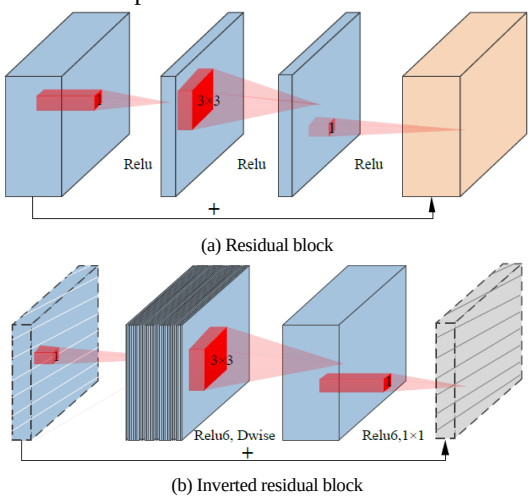


FIGURE 6.Depthwise Separable Convolution.

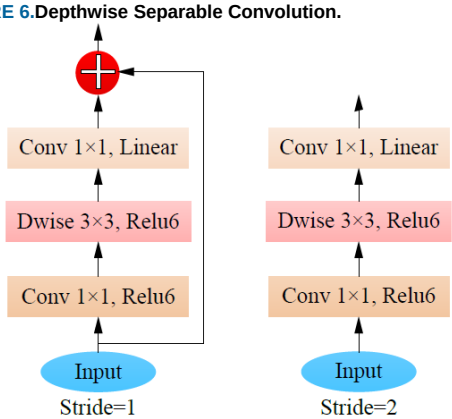


FIGURE 7.Inverted Residual Linear Bottleneck Structure.

4) LOSS FUNCTION IMPROVEMENT

In the natural environment, safflower filaments are harvested predominantly through unstructured planting. During the harvesting process, some filaments are obscured by leaves, capsules, and stems. The default Complete Intersection over Union (CIoU) in YOLOv5s loses effectiveness in handling overlapping objects because of the limited adjustment space for the position and shape of the bounding boxes. This limitation hinders the optimization of the loss function, which poses significant challenges for the recognition of safflower filaments. DIoU-NMS[32] can filter multiple detection boxes to reduce overlapping bounding boxes for targets of different scales and aspect ratios. Based on this, our study aimed to adopt the DIoU non-maximum

suppression method within the default CIoU loss function of YOLOv5s, effectively enhancing the detection performance of filaments in occluded environments. The optimized total loss function is given by Equation (10).

$$L_{CIoU(D-NMS)} = L_{CIoU} + S_i \tag{10}$$

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(\mathrm{b}, \mathrm{b}^{gt})}{c^2} + \alpha v \tag{11}$$

$$S_i = \begin{cases} S_i, & IoU - R_{DIoU}(M, B_i) < \varepsilon \\ 0, & IoU - R_{DIoU}(M, B_i) \geq \varepsilon \end{cases} \tag{12}$$

$$R_{DIoU} = \frac{\rho^2(\mathrm{b}, \mathrm{b}^{gt})}{c^2} \tag{13}$$

where: IoU - traditional regression loss, ρ - Euclidean distance, $\mathrm{b}, \mathrm{b}^{gt}$ - center points of predicted box A and target box B, c- the diagonal length of the smallest enclosing box covering both boxes, α - balance parameter, v - aspect ratio correction factor, M - the bounding box with the maximum confidence, B_i - other bounding boxes, ε - NMS threshold.

5) IMPROVED SAFFLOWER FILAMENT DETECTION MODEL FRAMEWORK BASED ON YOLOV5S

In the natural environment, intelligent picking of safflower filaments inevitably suffers from mis-detection and missed picking owing to factors such as filament occlusion, multiple target overlap, changes in illumination, and equipment performance. To address this issue, this study proposes an improved lightweight detection model for safflower filaments, YOLOv5s-MCD. It mainly replaces the backbone part of YOLOv5s with an improved MobileNetv2 lightweight network structure, embeds the CA module into the neck network layer to enhance the model's feature extraction capability, and uses the DIoU non-maximum suppression method to improve the default CIoU loss function of YOLOv5s. The overall network structure is shown in Figure 8.

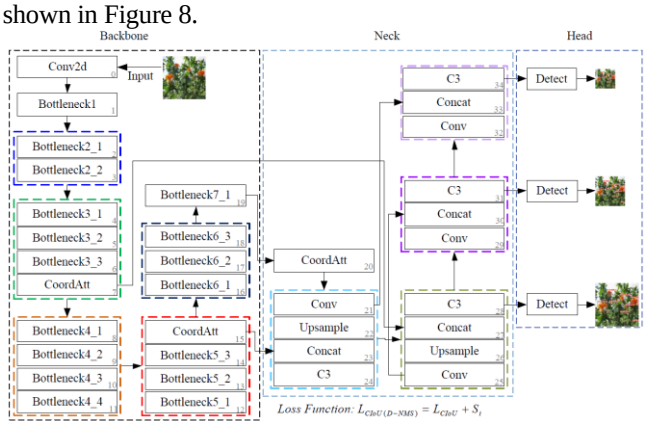


FIGURE 8.Overall framework of YOLOv5s-MCD safflower filament detection model.

D. EXPERIMENTAL PARAMETERS EVALUATION METRICS

1) EXPERIMENTAL PARAMETERS

The entire model training was conducted on a Windows 11 system, with computer hardware configured with 384 GB of memory and an AMD Ryzen Threadripper PRO 3975WX 32-Cores processor. The libraries required for the model configuration included Anaconda 3.8, Python 3.8, OpenCV, CUDA 11.6, and Pytorch 1.12 deep learning framework. A mosaic data augmentation method was employed during the model training process. A batch size of 16 was utilized for each processing unit, with Batch Normalization (BN) layers applied for regularization at each weight update. The initial learning rate was set as 0.01. The optimizer used was a Stochastic Gradient Descent (SGD) with a momentum factor of 0.937. Training spanned 300 epochs. After the completion of training, the weights with the best accuracy and the final training weights were saved. Optimal weights were used to evaluate the safflower filament test dataset and obtain the detection results.

2) EVALUATION METRICS

Multiple models were applied to the same test dataset in order to assess the performance of the improved model. When evaluating the performance of each model, six key metrics were considered [33-34]: Precision (P), Recall (R), Mean Average Precision (mAP), F1 Score, model size and detection response time. Among these metrics, the mAP and detection speed are particularly important for the model because they jointly determine the model's performance in terms of accuracy and speed. The formulas for calculating the precision, recall, F1, and mean average precision are as follows:

$$P = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \times 100\% \tag{14}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \times 100\% \tag{15}$$

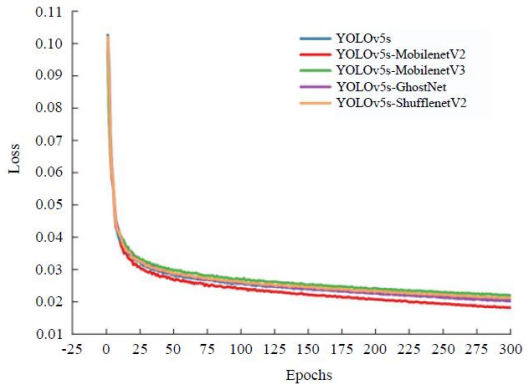
$$F1 = \frac{2 \times P \times R}{P + R} \times 100\% \tag{16}$$

$$\mathrm{mAP} = \frac{\sum_{i=1}^{N} \int_0^1 P(R)\mathrm{d}R}{N} \times 100\% \tag{17}$$

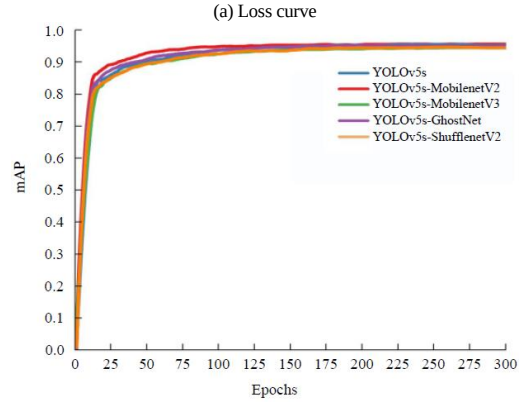
where, TP stands for true positive, FP stands for false positive, FN stands for false negative. TP represents the number of samples that the model predicts as positive and is actually positive, FP represents the number of samples that the model predicts as positive but is actually negative, FN represents the number of samples that the model predicts as negative but is actually positive, AP represents the average precision for a single category, and N is the number of categories to be detected. In this study, only blooming and withered stamens were detected; therefore, N=2.

III. Results and Discussion

A. MODEL TRAINING



(a) Loss curve



(b) mAP curve

FIGURE 9.Loss and mAP curves during the training process.

This study is based on the YOLOv5s model with different lightweight backbone networks to train on the safflower filament image dataset, as shown in Figure 9. Compared to other network models, the improved model YOLOv5s-MCD has the fastest convergence speed of the loss curve, the largest increase in the mean average precision (mAP), and when trained for up to 300 epochs, the loss value tends to be stable, fluctuating around 0.02, with an mAP value reaching 95.6%, and the model size is only 6.01MB. All performance parameters are better than those of the other models and are more suitable for deployment applications on low-computing power platforms in safflower filament fields. Therefore, this study used the best model output after 300 rounds of training as the recognition model for intelligent safflower filament harvesting.

B. ABLATION STUDY RESULTS ANALYSIS OF THE IMPROVED YOLOV5S-MCD MODEL

1) COMPARATIVE RESULTS OF EXPERIMENTS BASED ON DIFFERENT NETWORK STRUCTURES OF YOLOV5S

To evaluate the detection efficiency of the safflower filament based on the YOLOv5s model, we trained different backbone network structures: MobileNetV2, MobileNetV3, ShuffleNetV2, and GhostNet using the same improvement methods on a unified test dataset. The test results are listed in Table 1. According to the training results, the model obtained by improving MobileNetV2 to replace the original backbone network performed better than the other three backbone network variants. Specifically, although YOLOv5s-M (YOLOv5s-MobileNetV2) is slightly lower than the original YOLOv5s model in terms of Precision (P), Recall (R), F1 Score, and mean Average Precision (mAP), the model size has been reduced from 13.7 MB to 5.98 MB, a reduction of 56.4%, which is more advantageous for deployment on low-computing platforms for detecting Safflower filaments. Compared with the other three models, the improved YOLOv5s-M model not only has an advantage in size but also slightly outperforms it in over 90% of the indicators. In summary, the proposed YOLOv5s-M improved model significantly achieves a lightweight model while ensuring detection speed and accuracy. This indicates that the model has outstanding comprehensive performance and is particularly suitable for the rapid and accurate detection of small targets, such as safflower filaments.

TABLE 1
PERFORMANCE COMPARISON OF DIFFERENT MODELS BASED ON YOLOV5S

Model	P/%	R/%	F1/%	mAP@.5/%	Model Size/MB
YOLOv5s	93.8%	91.8%	92.8%	95.6%	13.7
YOLOv5s+ghost	92.3%	91.0%	91.6%	95.1	7.44
YOLOv5s+mobilenetv3	93.0%	89.7%	91.3%	94.5%	7.11
YOLOv5s+shufflenetv2	91.9%	90.1%	91.0%	94.7%	7.63
YOLOv5s+mobilenetv2	92.4%	91.5%	91.9%	95.5%	5.98

2) COMPARISON OF EXPERIMENTAL RESULTS OF DIFFERENT ATTENTION MECHANISMS ON YOLOV5S-M MODEL

To evaluate the effect of different attention mechanisms on the performance of the improved YOLOv5s-M model, performance tests were conducted on the improved YOLOv5s-M model using four types of attention mechanisms: CBAM, SE, CA, and ECA. Each enhanced model was trained under the same conditions for 300 epochs and was evaluated using the same test dataset. The performance test results for the four attention mechanisms are presented in Table 2. By comparing the experimental results, it can be found that the model with the CA attention mechanism has an accuracy of 95.2% and 92.7% for the Safflower-B and Safflower-D classes, respectively, which is an increase of 1.1 and 1.9 percentage points compared with the original YOLOv5s-M model without any attention mechanism. It is also an increase of 2.3 and 2.0 percentage points compared to the model with the CBAM attention mechanism, an increase of 1.2 and 1.4 percentage points compared to the model with the SE attention mechanism, and an increase of 1.5 and 1.2 percentage points compared to the model with the ECA attention mechanism. Moreover, the detection time for a single image is the shortest among the five models, at only 3.2 ms. The recognition of safflower filaments mainly occurs on a fast-moving platform in the

field, which requires high accuracy and speed. Based on the analysis of the experimental results, the improved YOLOv5s-MC (YOLOv5s-MobileNetV2-CA) model has incomparable advantages.

TABLE 2
PERFORMANCE COMPARISON OF YOLOV5S-M WITH DIFFERENT ATTENTION MECHANISMS ADDED

Model	Category	P/%	R/%	F1/%	mAP@.5/%	Detection time /ms
YOLOv5s-M	all	92.4%	91.5%	91.9%	95.5%	3.4
	Safflower-B	94.1%	93.6%	93.8%	97.2%	
	Safflower-D	90.8%	89.4%	90.1%	93.8%	
YOLOv5s-M-CBAM	all	91.8%	91.4%	91.6%	95.6%	3.3
	Safflower-B	92.9%	93.8%	93.3%	97.2%	
	Safflower-D	90.7%	88.9%	89.8%	93.9%	
YOLOv5s-M-SE	all	92.6%	90.3%	91.4%	95.4%	3.3
	Safflower-B	94.0%	92.8%	93.4%	97.3%	
	Safflower-D	91.3%	87.8%	89.5%	93.5%	
YOLOv5s-M-CA	all	93.9%	90.1%	92.0%	95.5%	3.2
	Safflower-B	95.2%	91.6%	93.4%	97.2%	
	Safflower-D	92.7%	88.6%	90.6%	93.8%	
YOLOv5s-M-ECA	all	92.6%	90.7%	91.6%	95.2%	3.4
	Safflower-B	93.7%	93.4%	93.5%	97.2%	
	Safflower-D	91.5%	87.9%	89.7%	93.2%	

3) COMPARISON OF EXPERIMENTAL RESULTS OF DIFFERENT LOSS FUNCTION ON YOLOV5S-MC MODEL

To further improve the overall comprehensive recognition performance of the final improved model, ablation experiments were conducted on the loss function of YOLOv5s-MC. The comparative experimental results are shown in Table 3. The analysis showed that the CIoU loss function with the DIoU non-maximum suppression method proposed in this study achieved a comprehensive mAP of 95.6% for the recognition of safflower filaments, 97.4% for recognizing blooming filaments, and 93.8% for recognizing wilting filaments. Compared to the other loss functions of YOLOv5s-MC, CIoU(D) demonstrated a better comprehensive recognition performance for safflower filaments. Although the mAP value for Safflower-D is slightly lower by 0.1% compared to the original YOLOv5s model, the model size has been reduced from 13.7 MB to 6.01 MB and other mAP values remain balanced. This indicates that the YOLOv5s-MCD (YOLOv5s-MobileNetV2-CA-CIoU(D)) model proposed in this study is more suitable for applications in intelligent safflower filament harvesting systems in the field.

TABLE 3
PERFORMANCE COMPARISON OF DIFFERENT LOSS FUNCTIONS

Model	Model Size/MB	Loss Function	Category	mAP@.5/%
YOLOv5s	13.7	CIoU	all	95.6%
			Safflower-B	97.4%
			Safflower-D	93.9%
YOLOv5s-MC	6.01	CIoU	all	95.5%
			Safflower-B	97.2%
			Safflower-D	93.8%
		CIoU(D)	all	**95.6%**
			Safflower-B	**97.4%**
			Safflower-D	**93.8%**
		GIoU	all	95.2%
			Safflower-B	97.1%
			Safflower-D	93.3%
		SIoU	all	95.3%
			Safflower-B	97.2%
			Safflower-D	93.3%
		EIoU	all	95.3%
			Safflower-B	97.4%
			Safflower-D	93.2%

4) THE EXPERIMENT RESULTS COMPARING YOLOV5S-MCD MODEL WITH THE BASELINE MODEL

To ascertain the superior processing performance of the proposed YOLOv5s-MCD model in safflower filament recognition, it was juxtaposed against the base models YOLOV5s, YOLOv8s, and YOLOv9 in terms of the detection average precision, model size, and detection time. The experimental results are presented in Table 4. As shown in Table 4, the YOLOv5s-MCD model, an enhancement of YOLOv5s, exhibits a smaller model size and detection time, showing heightened sensitivity in recognizing the safflower filaments dataset, with an average precision on par with YOLOv5s. Conversely, YOLOv8s and YOLOv9, despite being the latest algorithmic models, demonstrate slightly lower average detection precision than YOLOv5s-MCD and possess larger model volumes, making them less deployable on low-computing power mobile platforms. Consequently, the improved YOLOv5s-MCD model presents a more

reliable value proposition for deployment in safflower filament field automation harvesting equipment.

TABLE 4
COMPARISON OF PERFORMANCE BETWEEN YOLOV5S-MCD AND VARIOUS BASELINE MODELS

Model	Category	mAP@.5/%	Model Size/MB	Detection time /ms
YOLOv5s	Safflower-B	97.4%	13.7	3.2
	Safflower-D	93.9%		
YOLOv8s	Safflower-B	96.6%	21.4	3.4
	Safflower-D	92.9%		
YOLOv9	Safflower-B	96.9%	135.0	14.5
	Safflower-D	93.1%		
YOLOv5s-MCD	**Safflower-B**	**97.4%**	**6.01**	**3.2**
	Safflower-D	**93.8%**		

C. CONFIDENCE COMPARISON

Three randomly selected images (A1, A2, A3) were tested using the YOLOv5s and YOLOv5s-MCD models, and the results are shown in Figure 10, with the confidence levels of safflower filament detection presented in Table 5. The YOLOv5s-MCD model detected 5, 2, and 3 safflower filaments in these three images, respectively, whereas the YOLOv5s model detected 5, 2, and 2 filaments, with missed detections occurring in backlit conditions. Compared with the YOLOv5s model, the YOLOv5s-MCD model occupies less memory and has increased confidence in detecting safflower filaments, resulting in better detection outcomes.

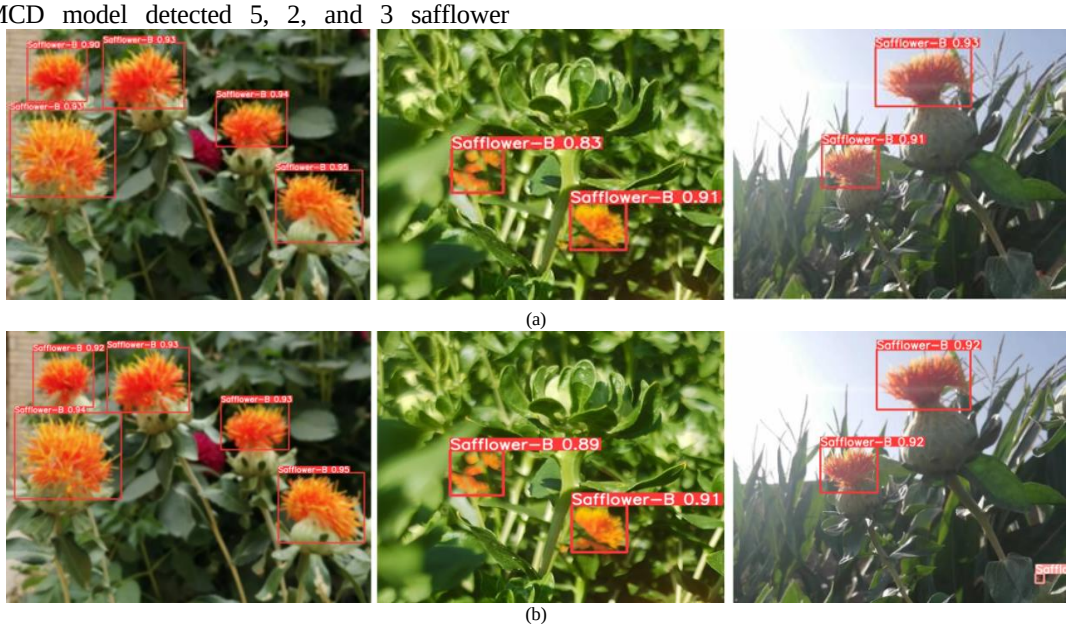


FIGURE 10.Comparison of Detection Results between YOLOv5s and YOLOv5s-MCD Models. (a) YOLOv5s; (b) YOLOv5s-MCD.

TABLE 5
CONFIDENCE COMPARISON RESULTS

Model	Model Size/MB	Image Number	Number of filaments	Confidence
YOLOv5s	13.7	A1	5	0.95、0.94、0.93、0.90、0.93
		A2	2	0.83、0.91
		A3	2	0.93、0.91
YOLOv5s-MCD	6.01	A1	5	0.95、0.93、0.93、0.92、0.94
		A2	2	0.89、0.91
		A3	3	0.92、0.92

D. HEATMAP VISUALIZATION ANALYSIS

Heatmap visualization analysis is a powerful visual aid that can intuitively reveal key areas in images and reflect the focus of the attention of models during object detection tasks. Grad-CAM is a widely used gradient-based visualization technique [35] that generates heatmaps that can be overlaid with original images to highlight the areas considered most critical by the model during prediction. The heatmap visualization results of YOLOv5s and YOLOv5s-MCD on the target of safflower filaments in natural scenes are presented in Figure 11. Through a comparative analysis, it can be observed that the heatmap generated by the YOLOv5s-MCD model is more concentrated and complete, showing a higher degree of attention to the safflower filament target area. Compared with the original YOLOv5s model, the improved model has a better detection performance for safflower filaments in their natural growing environment.

FIGURE 11.Visualization analysis of heatmap. (a) Original Image; (b) YOLOv5s; (c) YOLOv5s-MCD.

E. COMPARISON OF DECETION RESULTS IN NATURAL ENVIRONMENTS

To validate the robustness of the YOLOv5s-MCD model in natural environments, five datasets were constructed for different scenarios: unobstructed, obstructed, backlighting, shaking, and wide angle. Each scenario dataset comprised 50 images and was compared with the YOLOv5s model. The performance of the improved model was evaluated using the Precision (P), Recall (R), Mean Average Precision (mAP), and F1-score. The P-value measures the probability of correctly identifying the Safflower-B and Safflower-D filaments; a higher P-value indicates better accuracy in recognizing these filaments. The R-value assesses the probability of correctly identifying all filaments as filaments; a higher R-value indicates a lower probability of missed detection. The mAP measures the average precision across both categories, while the F1-score balances the relationship between P and R. These two metrics comprehensively evaluate the performance of the model in recognizing safflower filaments, with higher values indicating a higher recognition accuracy for Safflower-B and Safflower-D. The detection results are listed in Table 6, and the detection images are illustrated in Figure 12.

An analysis of the results from Table 6 and Figure 12 shows that under unobstructed natural conditions, the YOLOv5s-MCD model has a higher accuracy rate P, recall rate R, mean Average Precision (mAP), and F1 score by 5.7, 0.5, 4.4, and 2.9 percentage, respectively , compared to the YOLOv5s model. Additionally, the overall confidence level is higher than that of the YOLOv5s model, which exhibits instances of missed detection marked with blue arrows in the figure. When safflower filaments are occluded, the YOLOv5s-MCD model's recall rate R, mean Average Precision (mAP), and F1 score surpass those of the YOLOv5s model by 6.6, 0.7, and 0.4 percentage, respectively, although its accuracy rate P is slightly lower than that of the YOLOv5s model. The YOLOv5s model mistakenly identifies two different categories of the same filament, as indicated by yellow arrows in the figure. Under backlighting conditions, the YOLOv5s-MCD model achieved an accuracy rate P, mean Average Precision (mAP), and F1 scores of 91.2%, 89.7%, and 83.8% respectively, which were 17.1, 3.3, and 3.9 percentage points higher than those of the YOLOv5s. The overall confidence level of the YOLOv5s-MCD model was higher than that of the original YOLOv5s model, which still suffered from missed detections, as indicated by the yellow and blue arrows in the figure. In scenarios with shaking, both models perform poorly in recognition, but the improved model still outperforms the original YOLOv5s model with a higher accuracy rate P, mean Average Precision (mAP), and F1 score by 13.1, 3.4, and 3.1 percentage, respectively, and the overall confidence levels are slightly better. However, the original model has detection omissions, as indicated by the blue arrows in the figure. For safflower filaments observed at wide angles, the improved YOLOv5s-MCD model's accuracy rate P and mean Average Precision (mAP) were 8.7 and 1.0 percentage points higher, respectively than those of the original model. Although the recall rate R is slightly lower than that of the original YOLOv5s model, the latter has instances of missed and incorrect detections, which are marked with blue and yellow arrows in the figure. Because actual picking operations mainly focus on close-range small angles, and

wide angles are only used for initial preparation, these issues do not affect the overall recognition performance.

TABLE 6
THE DETECTION RESULTS OF YOLOV5S AND YOLOV5S-MCD MODELS IN NATURAL ENVIRONMENTS

Different Scenarios	Evaluation Criteria	YOLOv5s	YOLOv5s-MCD
Unobstructed	P	88.5%	94.2%
	R	83.1%	83.6%
	mAP@.5/%	87.1%	91.5%
	F1	85.7%	88.6%
Obstructed	P	85.0%	78.7%
	R	76.1%	82.7%
	mAP@.5/%	81.2%	81.9%
	F1	80.3%	80.7%
Backlight	P	74.1%	91.2%
	R	86.8%	77.5%
	mAP@.5/%	86.4%	89.7%
	F1	79.9%	83.8%
Shaking	P	35.1%	48.2%
	R	64.3%	48.9%
	mAP@.5/%	43.3%	46.7%
	F1	45.4%	48.5%
Wide angle	P	68.0%	76.7%
	R	83.5%	73.4%
	mAP@.5/%	77.7%	78.7%
	F1	75.0%	75.0%

(e)

FIGURE 12.YOLOv5s and YOLOv5s-MCD detection performance in natural environments. (a) Unobstructed; (b) Obstructed; (c) Backlight; (d) Shaking; (e) Wide angle.

Based on the analysis of the experimental results, the baseline model YOLOv5 exhibited issues of missing and incorrect detection of safflower filaments in natural scenes. These problems are particularly prominent in small, distant, and severely occluded filaments. The primary reason for these shortcomings is the limited capacity of CNNs for global feature extraction within YOLOv5, which hinders its ability to capture comprehensive feature information. However, the improved YOLOv5s-MCD model, which incorporates CA for dynamically adjusting the global attention weights, demonstrates higher accuracy and greater robustness. In additionally, it significantly reduces memory usage compared to the original YOLOv5s model and consistently shows reliable detection performance across various natural working environments.

F. DISCUSSION

Based on the experimental results, this study demonstrates that the proposed YOLOv5s-MCD model maintains the detection accuracy while significantly reducing the model weight. This enhancement addresses the issues of low detection accuracy and large model size in natural scenarios, making it more suitable for deployment on low-power mobile platforms. Furthermore, the improved model outperformed the state-of-the-art YOLOv9 algorithm by achieving a smaller model size and higher detection accuracy, particularly on the safflower filament dataset. However, deploying YOLOv5s-MCD on field robots for safflower filament detection requires the consideration of hardware compatibility and integration with the robot control system. To ensure hardware compatibility, it is essential to ensure that the hardware and drivers are compatible with the YOLO framework (PyTorch), considering the device's computational power and memory constraints. The NVIDIA Jetson Orin NX 16GB, with its 100TOPS computational power, can be selected for edge deployment to enhance the real-time identification performance of YOLOv5s-MCD in the field. In terms of integration with the robot control system, the control system should be modularized with the visual control system, appropriate data interfaces and communication protocols should be designed, and a real-time feedback mechanism should be established to adjust the robot's positioning and picking actions based on the YOLOv5s-MCD detection results. This is the focus of the next phase of research.

In future research endeavors, the focus will be on determining the spatial picking points of safflower filaments through the integration of RGB and depth information, the combination of morphological characteristics and algorithms, and the lightweight design of end effectors. This approach aims to further optimize localization algorithms suited for the visual picking of safflower filaments. Additionally, it is imperative to investigate multi-arm coordination and the motion control strategies of end effectors to address the challenges posed by occlusion of filaments or the invisibility of flowers and fruits, which can result in the failure of picking point estimation.

IV. CONCLUSION

(1) To address the problems of low recognition accuracy and large model volume of existing safflower filament recognition algorithms, safflower filament dataset was developed in natural environmental conditions(the dataset has been published in the https://gitcode.net/m0_60172526/yolov5-mcd), a real-time recognition algorithm for safflower filament picking based on YOLOv5s aimed at a lightweight network structure, YOLOv5s-MCD, was proposed. It realizes the picking recognition of safflower filaments in different natural environments, thus providing visual guidance for the robotic arm to actively adjust its posture to avoid obstructions by fruit balls and leaves during filament picking.

(2) Based on the YOLOv5s backbone network, it has been replaced with an improved lightweight network structure, Mobilenetv2, which achieves a lightweight improvement over the original backbone network. The CA attention mechanism is embedded in the Neck network layer, which enhances fusion processing performance by extracting different levels of features from the improved backbone network. The original YOLOv5s default CIoU non-maximum suppression method was improved, effectively increasing the detection accuracy and convergence speed of the filament model.

(3) The proposed YOLOv5s-MCD recognition algorithm can effectively recognize harvestable and non-harvestable safflower filaments. By comparing different lightweight backbone networks, this algorithm was found to have the smallest model volume and the best recognition performance. The test set experimental results showed that the algorithm reduced the model volume to 55.5% of the original YOLOv5s model, with the mean average precision of

blooming safflower filament recognition reaching 97.4%, and the average recognition time per image was 3.2 ms.

(4) The improved YOLOv5s-MCD model was compared and analyzed with the YOLOv5s model under unobstructed, obstructed, backlighting, shaking, and wide-angle conditions. In the case of unobstructed filaments, P, R, mAP, and F1 scores were 5.7, 0.5, 4.4, and 2.9 percentage points higher than the YOLOv5s model; in the case of obstructed filaments, R, mAP, and F1 scores were 6.6, 0.7, and 0.4 percentage points higher; in the case of backlighting, P, mAP, and F1 scores were 17.1, 3.3, and 3.9 percentage points higher; in the case of shaking, P, mAP, and F1 scores were 13.1, 3.4, and 3.1 percentage points higher, with overall higher confidence; and in the case of wide-angle, P and mAP were 8.7 and 1.0 percentage points higher. This model improves the picking recognition performance of safflower filaments in different natural environments, providing an important reference for the development of safflower picking robots.

REFERENCES

[1] National Pharmacopoeia Commission, *Pharmacopoeia of the People's Republic of China*. Beijing, China: China Medical Science and Technology Press, 2020, pp. 232-233.

[2] Y. Ge, L. Zhang, D. Han, J. Chen, and W. Fu, "Current state and development trend of the mechanical harvesting on saffron filaments," *J. Agric. Mech. Res*., vol. 36, no. 11, pp. 265-268, Nov. 2014, doi: 10.13427/j.cnki.njyi.2014.11.062.

[3] Y. Zhou, J. Guo, X. Ma, X. Fan, Y. Chen, and M. Lin, "Research on current situation and development countermeasures of Xinjiang safflower production," *J. Anhui Agric. Sci*., vol. 49, no. 19, pp. 199-201+217, Oct. 2021.

[4] W. Cao, H. Jiao, J. Liu, S. Yang, B. Chen, and W. Sun, "Design of safflower filament picking device based on TRIZ theory," *Trans. Chin. Soc. Agric. Mach*., vol. 49, no. 08, pp. 76-82, Aug. 2018.

[5] Z. Zhang, Z. Xing, M. Zhao, S. Yang, Q. Guo, R. Shi, and C. Zeng, "Detecting safflower filaments using an improved YOLOv3 under complex environments," *Trans. Chin. Soc. Agric. Eng*., vol. 39, no. 03, pp. 162-170, Mar. 2023.

[6] Z. Zhang, R. Shi, Z. Xing, Q. Guo, and C. Zeng, "Improved faster region-based convolutional neural networks(R-CNN) model based on split attention for the detection of safflower filaments in natural environments," *Agronomy*. vol. 13, no. 10, pp. 2596, Oct. 2023, doi: 10.3390/agronomy13102596.

[7] X. Wang, Y. Xu, J. Zhou, and J. Chen, "Safflower picking recognition in complex environments based on an improved YOLOv7," *Trans. Chin. Soc. Agric. Eng*., vol. 39, no. 06, pp. 169-176, Mar. 2023.

[8] F. Dong, H. Guo, J. Pan, and Y. Han, "Study on safflower image segmentation based on an improved color difference model," *For. Mach. Wood. Equip*., vol. 51, no. 08, pp. 68-74, Oct. 2023, doi: 10.13279/j.cnki.fmwe.20231019.004.

[9] X. Fan, J. Zhou, Y. Xu, K. Li, and D. Wen, "Identification and localization of weeds based on optimized Faster R-CNN in cotton seedling stage," *Trans. Chin. Soc. Agric*., vol. 52, no. 05, pp. 26-34, Mar. 2021.

[10] Y. Shang, Q. Zhang, and H. Song, "Application of deep learning using YOLOv5s to apple flower detection in natural scenes," *Trans. Chin. Soc. Agric. Eng*., vol. 38, no. 09, pp. 222-229, May. 2022.

[11] D. Wu, S. Lv, M. Jiang, and H. Song, "Using channel pruning-based YOLO v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments," *Comput. Electron. Agric*.**,** vol. 178, pp. 105742, Nov. 2020, doi: 10.1016/j.compag.2020.105742.

[12] C. Chen, J. Lu, M. Zhou, J. Yi, M. Liao, and Z. Gao, "A YOLOv3-based computer vision system for identification of tea buds and the picking point," *Comput. Electron. Agric*., vol. 198, pp. 107116, Jun. 2022, doi: 10.1016/j.compag.2022.107116.

[13] J. Li, J. Li, X. Zhao, X. Su, and W. Wu, Lightweight detection networks for tea bud on complex agricultural environment via improved YOLOv4. *Comput. Electron. Agric*., vol. 211, pp. 107955, Jun. 2023, doi: 10.1016/j.compag.2023.107955.

[14] H. Miao, Z. Li, and J. Wu, "Lightweight maturity detection of cherry tomato based on improved YOLOv7," *Trans. Chin. Soc. Agric. Mach*., vol. 54, no. 10, pp. 225-233, Oct. 2023.

[15] C. Han, C. Yan, S. Qiu, Y. Xu, B. Hu, and H. Mao, "Design and experiment of double disc cotton topping device based on machine vision," *Trans. Chin. Soc. Agric. Mach*., vol. 54, no. 05, pp. 36-46, Mar. 2023.

[16] Y. Shang, X. Xu, Y. Jiao, Z. Wang, Z. Hua, and H. Song, "Using lightweight deep learning algorithm for real-time detection of apple flowers in natural environments," *Comput. Electron. Agric*., vol. 207, pp. 107765, Mar. 2023, doi: 10.1016/j.compag.2023.107765.

[17] L. Shuai, J. Mu, X. Jiang, P. Chen, B. Zhang, H. Li, Y. Wang, and Z. Li, "An improved YOLOv5-based method for multi-species tea shoot detection and picking point location in complex backgrounds," *Biosyst. Eng*., vol. 231, pp. 117-132, Jun. 2023, doi: 10.1016/j.biosystemseng.2023.06.007.

[18] S. Li, J. Bian, K. Li, and H. Ren, Identification and height localization of sugarcane tip bifurcation points in complex environments based on improved YOLOv5s. *Trans. Chin. Soc. Agric., Mach*. vol. 54, no. 11, pp. 247-258, Sep. 2023.

[19] YOLOv5. https://github.com/ultralytics/yolov5. 1 October 2023.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 37, no. 09, pp. 1904-1916, Sep. 2015, doi: 10.1109/tpami.2015.2389824.

[21] C. Szegedy, S. Loffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," In *Proc. AAAI Artif. Intell*., San Francisco, California, USA, 2017, pp. 4278-4284.

[22] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and Excitation networks," In *Proc. IEEE Comput. Vision & Pattern Recognit*., Salt Lake City, UT, USA, 2018, pp.7132-7141.

[23] S. Woo, J. Park, J.-Y. Lee, and I.S. Kweon, "CBAM: Convolutional Block Attention Module," In *Proc. ECCV*., Salt Lake City, UT, USA, 2018, pp. 3-19.

[24] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, (2020). "ECA-Net: Efficient channel attention for deep convolutional neural networks," In *Proc. IEEE/CVF Comput. Vision & Pattern Recognit.*, Scotland, UK, 2020, pp. 11534-11542.

[25] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," In *Proc. IEEE Comput. Vision &Pattern Recognit*., Seattle, WA, USA, 2021, pp.13708-13717.

[26] J. Duan, Z. Wang, X. Zou, H. Yuan, G. Huang, and Z. Yang, "Recognition of bananas to locate bottom fruit axis using improved YOLOv5," *Trans. Chin. Soc. Agric. Eng*., vol. 38, no. 19, pp. 122-130, Oct. 2022.

[27] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," In *Proc. IEEE/CVF Comput. Vision & Pattern Recognit*., Salt Lake City, UT, USA, 2018, pp.4510-4520.

[28] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv* 2017, arXiv:1704.04861.

[29] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," In *Proc. AAAI Artif. Intell*., New York, USA, 2020, pp. 12993-13000.

[30] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, You Only Look once: Unified, Real-Time Object Detection. In *Proc. IEEE Comput. Vision & Pattern Recognit*., Las Vegas, NV, USA, 2016, pp.779-788.

[31] A. Lu, L. Ma, H. Cui, J. Liu, and Q. Ma, "Instance segmentation of lotus pods and stalks in unstructured planting environment based on improved YOLOv5," *Agriculture*, vol. 13, no. 08, pp. 1568, Aug. 2023. doi: 10.3390/agriculture13081568.

[32] A. Howard, M. Sandler, G. Chu, L. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, et al. "Searching for Mobilenetv3," In *Proc*. *IEEE/CVF Comput. Vision (ICCV)*, Seoul, Korea, 2019, pp.1314-1324.

[33] N. Ma, X. Zhang, H. Zheng, and J. Sun, "Shufflenet V2: Practical guidelines for efficient cnn architecture design," In *Proc*. *ECCV*., Munich, Germany, 2018, pp. 116-131.

BANGBANG CHEN obtained his Master's degree in Agricultural Machinery Engineering from Shihezi University in 2018. He is currently pursuing a Doctoral degree at Xi'an Technological University, with his research primarily focusing on the structural design of intelligent agricultural machinery equipment, machine vision inspection technology, and agricultural robot harvesting technology.

FENG DING , obtained his Ph.D. in Mechanical Engineering from Xi'an Jiaotong University, is a professor and doctoral supervisor at Xi'an Technological University. He serves as a peer review expert for the National Natural Science Foundation of China projects, and as an expert evaluator for Shaanxi Province's science and technology plan projects and scientific and technological awards. He has also been a visiting scholar and professor in the Department of Mechanical and Industrial Engineering at Ryerson University in Toronto, Canada. His main research areas include reliability assessment and optimization design, machine vision inspection technology, and digital integrated manufacturing technology.

BAOJIAN MA obtained his Ph.D. in Agricultural Engineering from Zhejiang University in 2021. He is currently a professor and master's supervisor at Xinjiang Institute of Technology. His main research interests are intelligent harvesting technology for agricultural robots, 3D point cloud technology, and digital integrated manufacturing technology.

XIANGDONG LIU obtained his Ph.D. in Agricultural Engineering from Shenyang Agricultural University in 2015. He is currently the Dean of the School of Mechanical and Electrical Engineering at Xinjiang Institute of Technology, a professor, and a master's supervisor. His main research areas are the design and development of intelligent agricultural machinery equipment, hydraulic and pneumatic control technology, and the application technology of agricultural robots.

SHANPING NING obtained his Master of Engineering degree from Jiangsu University of Science and Technology in 2015. He is currently pursuing a Ph.D. at Xi'an Technological University, with research interests mainly in railway signal processing, intelligent fault diagnosis, and machine vision inspection technology.

[34] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," In *Proc*. *IEEE/CVF Comput. Vision & Pattern Recognit*., Seattle, WA, USA, 2020, pp.1580-1589.

[35] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis*., vol. 128, pp. 336-359, Feb. 2020. doi: 10.1007/s11263-019-01228-7.