Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 11, No. 1, 2025

Research on DETR-based Weed Detection Algorithm

Wenbing Liao [1] and Wenwen Li [2]

[1] School of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin Jilin 132022, China
[2] School of Mechanical and Control Engineering, Baicheng Normal College, Baicheng, Jilin 137000, China

Abstract: To address the problem of complex field conditions and high similarity between corn seedlings and weeds, this study proposes an improved DETR (Detection Transformer) model for weed detection in corn fields, which uses the CBAM convolutional attention mechanism in the DETR model, and uses a focal loss function instead of the traditional cross-entropy loss function to balance the number of positive and negative class samples of the data. Compared with the original model, the mAP (Mean Average precision) of the improved model was increased by 1.12% to 91.56%.

Keywords: Transformer; CBAM; Focal Point Loss; Weed Detection.

1. Introduction

Weed infestation is one of the three major biological disasters in agricultural production, which leads to the loss of world food production equivalent to about 1 billion people's rations each year and causes direct economic losses of about 120 billion dollars [1]. China is also one of the country's most seriously afflicted by the weed problem, and the invasion of weeds will make the crop yield decline, while the traditional weed control methods are no longer suitable for the current environment [2]. With the rapid development of computer vision and artificial intelligence technology, weed detection methods based on image processing and deep learning have gradually become a research hotspot. By using advanced image processing algorithms and deep learning models, weeds in agricultural fields can be identified and classified, laying the algorithmic foundation for agricultural weeding equipment to realize intelligent and automated, and realizing intelligent agricultural management [3,4].

In the early days, due to the influence of computer arithmetic power, deep learning methods were not popularized, and machine vision-based recognition methods were widely used in crop and weed recognition. Mathanker et al[5] extracted multiple features including color and texture in oilseed rape and wheat crops and used AdaBoost and Support Vector Machines to perform automatic weed recognition. Tian Ronghui et al [6] achieved accurate recognition of overlapping leaves and weeds based on image chunking and reconstruction combined with support vector machine.

Since traditional image algorithms mainly rely on manually designed feature extractors and lack adaptive learning ability, the generalization ability and robustness are relatively poor, while with the improvement of computational resources and the availability of large amounts of data, deep learning has attracted extensive attention from researchers, and detection methods based on convolutional neural networks have made significant progress. Zhong Bin et al [7] proposed a pasture weed detection based on an improved DINO detection network, which enhanced the extraction of features by introducing an attention mechanism to achieve better recognition accuracy. Jin X et al [8] used three convolutional neural networks (DenseNet, EfficientNet, ResNet) to detect weeds in individuals growing in bermudagrass stratum weed species as well as herbicide susceptible weeds. Moazzam S I et al [9] used a new two-stage approach to improve the classification accuracy of crop weeds. In the first stage, a binary pixel-level classifier was developed to segment the background and vegetation; in the second stage, a three-class pixel-level classifier was devised to classify the background, weeds, and tobacco. Jinyang Le et al [10] proposed a weed detection model YOLOV7-FWeed based on the improved YOLOV7, which used F-ReLU as the activation function of the convolutional module and added the MaxPool Multihead Self-Attention (M-MHSA) module in order to improve the accuracy of weed recognition. Zhang Packing [11] implemented weed detection using YOLOV5s.Ong P et al [12] used convolutional neural network (CNN) to detect weeds in cabbage fields using images acquired by UAVs and compared the performance of random forest (RF) and CNN and concluded that the overall accuracy of CNN is higher than that of random forest. Khan S D et al [13] based on encoder-decoder architecture designed A new deep learning architecture was designed, where the encoder part effectively combines a dense absorbing network and a spatial pyramid pooling module for multi-scale feature extraction; the decoder part contains a deconvolutional layer and an attention unit, which helps to recover the spatial information and improves the ability to accurately locate weeds and crops in the image.

The above detection algorithms have achieved better results in weed detection, but there are still some limitations, the weeds in the field are complex, there is occlusion, blurring and other conditions, Transformer is better at dealing with global information compared to convolutional networks, so this study uses the Transformer-based model for the weed detection task in order to achieve better results, and to provide a smart weed control equipment with a theoretical basis.

2. DETR Network Model

DETR (Detection Transformer) [14] is an end-to-end target detection network based on Transformer [15] proposed by Facebook team. Traditional detection methods typically use a two-stage process, where candidate frames are first generated, and then classification and bounding box regression is performed on these candidate frames. Compared to convolution-based detection networks, DETR eliminates processes such as candidate frame generation and non-maximal suppression (NMS), transforming the target detection task into an ensemble prediction problem and

greatly simplifying the target detection process.

The model structure of DETR is shown in Fig. 1, which can be divided into the following three parts: first, the CNN backbone network for image feature extraction using ResNet network; second, the encoder and decoder based on Self-Attention Transformer; and third, the feed-forward network, FFN, which carries out the final detection. the original image is fed into the ResNet network to generate a feature map, and then undergoes a 1×1 convolution to convert the channel dimension of the ResNet output into a smaller dimension d to obtain a new feature map d×H×W. Transformer expects a sequence as input, so we collapse the spatial dimension of the new feature map into a single dimension to obtain d×HW, which is then passed as an input with the addition of positional encoding to the Transformer. the DETR decoder decodes N objects in parallel, and the N object queries are converted into one output embedding by the decoder. The feed-forward network decodes them independently as bounding box coordinates, predicting the normalized center coordinates, height and width of the bounding box, while the linear layer predicts the category labels using a softmax function. The DETR model exploits the self-attention of these embeddings and the encoder-decoder attention to reason globally over all the objects, being able to use the whole image as a context.

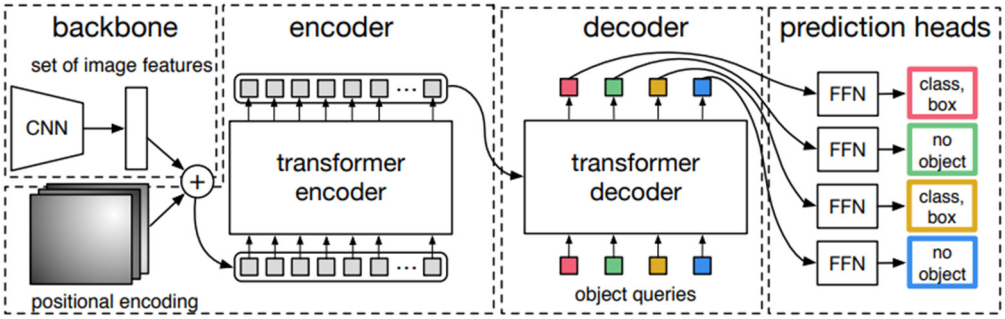


Fig 1. DETR model architecture

The DETR model uses an ensemble-based loss function that predicts a fixed-size ensemble of N bounding boxes that can be predicted for all objects simultaneously. N is usually set to be much larger than the actual number of objects of interest in the image, and requires an additional special class labeled "Ø" to indicate that no objects are detected within a slot. No objects are detected. this class acts similarly to the background class in standard object detection methods.

3. Model Algorithm Improvement

3.1. Focal Loss Function

In most cases the foreground region is smaller than the background region, so the number of anchors corresponding to the foreground target tends to be smaller than the number of anchors corresponding to the background target, and there is a serious classification imbalance problem in the foreground and background, and this imbalance problem affects the training process of the model. It leads to two problems: (1) training is inefficient because most places are simple negative examples, which do not help the model learning; (2) too many negative samples will dominate the model training and lead to lower model accuracy.

DETR infers a fixed-size set of N predictions in a single pass through the decoder, and DETR produces an optimal dichotomous match between predicted and real objects, optimizing object-specific losses. First, DETR performs a bipartite graph matching using an improved Hungarian algorithm, which matches the predicted set with the true set one by one, so that the matching loss is minimized, and a prediction result that minimizes the loss with the current true value is obtained by the Hungarian algorithm. After that, the loss function is calculated, which is a linear combination of classification loss and bounding box loss. The categorization loss is calculated from the cross-entropy of the categorized predicted value and the true value, and the bounding box loss is a linear combination of the L1 loss and the GIOU loss.

The standard cross-entropy loss function has the same contribution weight for each sample, and when there is an imbalance between positive and negative samples, the model training process may have a dominant effect, resulting in a weaker model for positive sample discrimination. The standard cross entropy is shown in equation (3):

$$CE(p,y) = \begin{cases} -log(p) & y = 1 \\ -log(1-p), & \text{otherwise} \end{cases} \quad (1)$$

where $y \in \{-1,1\}$, representing positive and negative samples, is the labeling probability predicted by the model, and if $p > 0.5$, it represents a positive sample, otherwise it is a negative sample. Define P_t by equation (2) and rewrite the loss function as $\mathrm{CE}(p_t) = -\log(p_t)$.

$$p_t = \begin{cases} p, & y = 1 \\ 1-p, & otherwise \end{cases} \quad (2)$$

In order to improve the effect of target detection, this paper replaces the cross-entropy loss function in the classification loss with the focus loss function [16] for calculation. Focus loss solves the problems of positive and negative sample extreme imbalance and difficult to classify sample learning by adjusting the weight of easy to classify samples, which makes the model pay more attention to difficult to classify samples. The focus loss formula is shown in (3):

$$FL_{p_t} = -\alpha_t(1-p_t)^{\gamma} log(p_t) \quad (3)$$

The definition of α_t is similar to that of p_t. In the formula for focus loss, α_t weights are introduced to improve the imbalance of positive and negative samples, and a modulation factor $(1-p_t)^{\gamma}$ is added to regulate the weights of the difficult and easy samples. when a border is misclassified, $(1-p_t)^{\gamma}$ is close to 1, and its loss is almost unaffected, and when p_t is close to 1, which indicates that it has better classification prediction and is a simple sample, $(1-p_t)^{\gamma}$ is close to 0, so its loss is moderated down. γ is a moderating factor, and the larger γ is, the lower the contribution of the simple sample loss will be.

3.2. CBAM Attention Module

CBAM [17], in order to emphasize features that are

meaningful along the two main dimensions, channel and space, applies channel and spatial attention modules sequentially, allowing each branch to learn what is important on the channel and spatial axes, respectively. The modular structure of CBAM is shown in Fig. 2.Given an intermediate feature map, the CBAM module sequentially infers the attention map along the two separate dimensions (channel and spatial), the graph, and then multiplies the attention graph by the input feature graph for adaptive feature refinement. The module can optimize the information flow process within the network by learning which information in the features needs to be emphasized or suppressed.

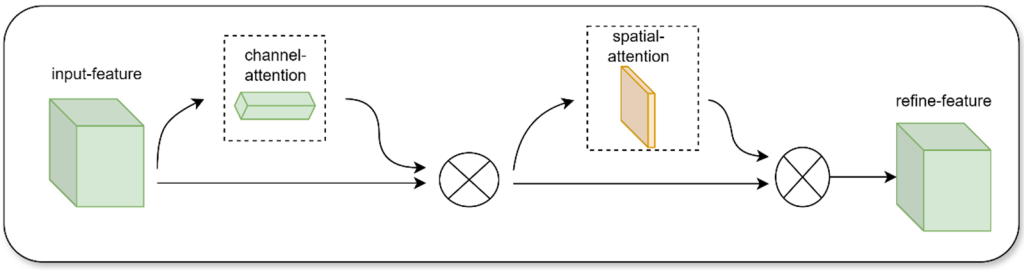


Fig 2. CBAM structure

The overall attention generation process can be summarized as shown in equations (4), (5):

$$F' = M_c(F) \otimes F \quad (4)$$

$$F'' = M_s(F') \otimes F' \quad (5)$$

F denotes the features input to the CBAM module, and M_c and M_s are the CBAM-derived 1D channel attention and 2D spatial attention, respectively, where $\otimes$ denotes multiplication by elements

Each channel of the feature map is considered as a feature detector and the channel attention is focused on the meaningful features of the given input image. To compute the channel attention efficiently, the spatial dimension of the input feature map is squeezed. The spatial attention map is generated by exploiting the spatial relationships of the features. Unlike channel attention, spatial attention concentrates on the effective information locations on the feature map and complements channel attention. To compute spatial attention, average pooling and maximum pooling operations are first applied along the channel axis and connected to generate a valid feature descriptor.

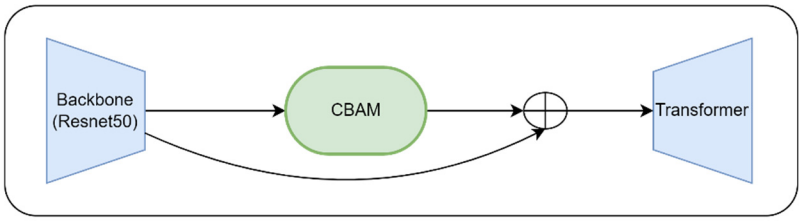


Fig 3. Model structure

The structure of the model with the addition of the CBAM module is shown in Fig. 3. After the original image is processed by the ResNet network, the features are refined using the CBAM attention module, which helps the network to focus on the regions and feature channels that are more important to the current task, thus improving the feature extraction ability of the network and the final recognition effect. At the same time, jump connections are added after CBAM to preserve the output features in the original network, and the information is more easily propagated to the later networks to avoid information loss.

4. Experimental Results and Analysis

4.1. Introduction to the Dataset

In this study, we used a dataset from the corn weed dataset that is publicly available online [18], and changed the dataset to have five categories, which were taken from the natural environment of corn seedling field, and images were collected under different soil background and sunlight conditions. The dataset contains maize seedlings with four common weeds such as prickly adelgid, sedge, quinoa and early morning glory, a total of 5 medium field plants, and the image size is 800 × 600.Meanwhile, in order to make the model learn a good characterization, the dataset is subjected to data enhancement. The diversity of data samples is improved by flipping, rotating, changing contrast, adding noise and other operations on the graphs to reduce the network overfitting phenomenon. The image enhancement result is shown in Fig. 4:

4.2. Assessment of Indicators

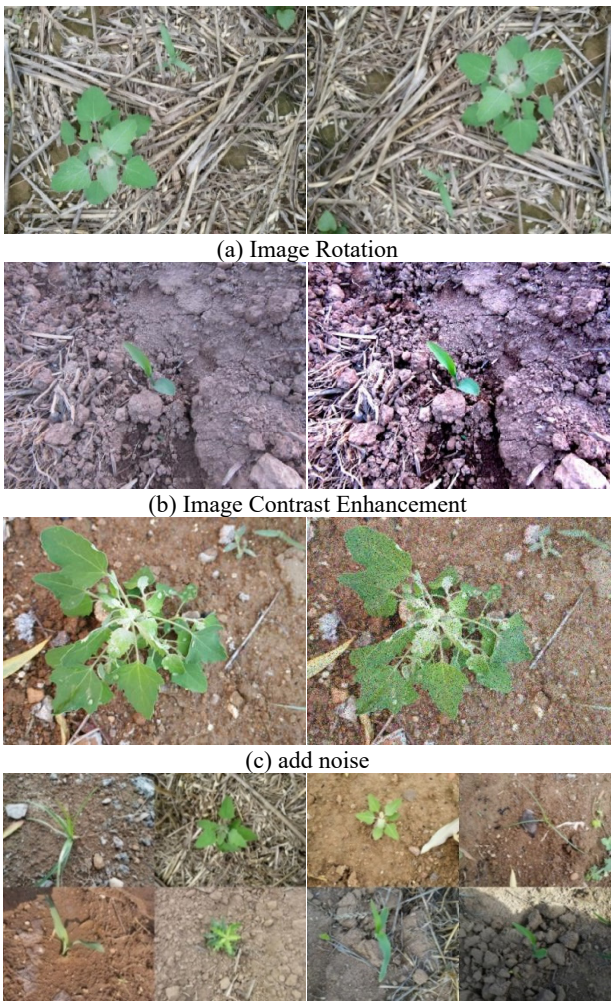
In this paper, we use mAP (Mean Average Precison) as the evaluation metrics of algorithm accuracy. mAP (Mean Average Precision) is one of the most common evaluation metrics in the task of target detection, which indicates the average of the detection precision (AP) of each category in the dataset, which can reflect the average accuracy performance of the model on all categories and measure the comprehensive performance of the model. The above evaluation metrics have formulas calculated as shown in (6)-(9):

$$\mathrm{P} = \frac{TP}{TP+FP} \quad (6)$$

$$\mathrm{R} = \frac{TP}{TP+FN} \quad (7)$$

$$\mathrm{AP} = \int_0^1 PdR \quad (8)$$

$$\mathrm{mAP} = \frac{\sum_{i=1}^{N} AP}{N} \quad (9)$$



(a) Image Rotation

(b) Image Contrast Enhancement

(c) add noise

(d) image stitching

Fig 4. Data enhancement

In the above equation: the number of positive samples predicted to be in the positive category, the number of negative samples predicted to be in the positive category, and the number of positive samples predicted to be in the negative category for TP, FP, and FN, respectively, P denotes the precision rate, i.e., the proportion of actual positive samples among the results predicted to be positive samples, and R stands for the recall rate, which denotes the proportion of positive samples correctly detected by the model as positive samples out of the positive samples used in the test set.

4.3. Results

The GPU used in this experiment is RTX3090, the python version is 3.8, and the deep learning framework is Pytorch, version 1.11.

During the training process, in order to evaluate the performance of the models more comprehensively, we performed a validation evaluation of the models after every 10 epochs and recorded the mAP change curves of each model in Figure 5. According to the mAP change curves of the models plotted in Fig. 5, we can see the impact of different algorithms on the model performance. By introducing the Convolutional Block Attention Mechanism (CBAM) and Focal Loss, the mAP of the models both achieved better results compared to the original model at the final convergence stage. In order to show the effect of the models in the training process more clearly, Table 1 lists the best mAP values achieved by each model in the evaluation process. Based on the contents of Table 1, it can be known that the incorporation of the CBAM attention mechanism and the focus loss function achieved the best test results in the dataset, with an improvement of 1.12% compared to the original model of DETR, and this enhancement suggests that the CBAM attention mechanism and the focus loss function play an obvious role in improving the model performance.

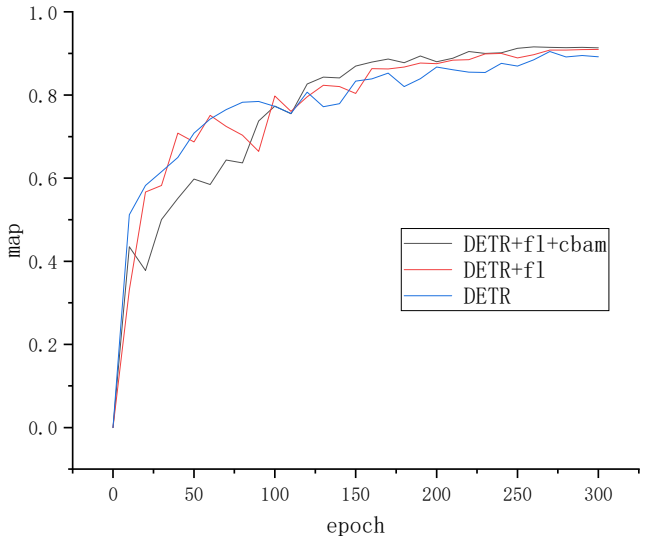
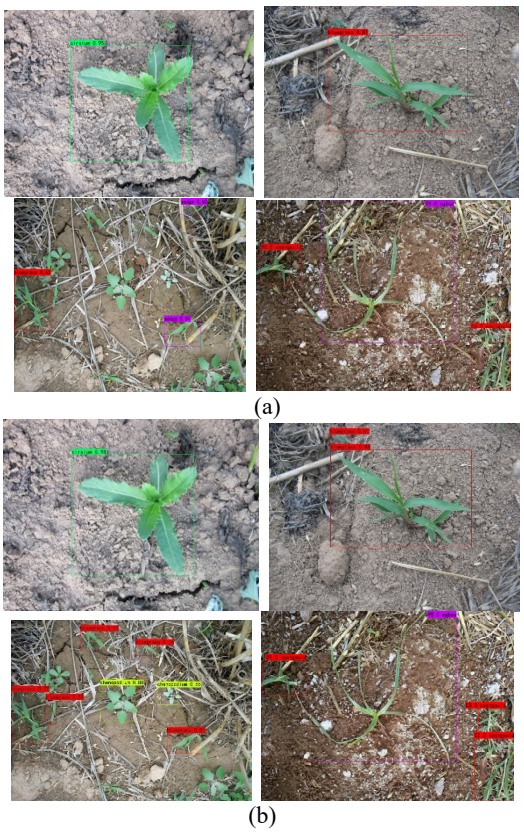


Fig 5. mAP change curve

Table 1. Results of DETR experiments

	Focal loss	CBAM	mAP
DETR			90.44
	√		90.96
	√	√	91.56



(a)

(b)

Fig 6. Model detection results

Fig. 6 shows the test results of the original DETR model and the improved model on the test set, and four images are selected to compare the performance of the models. Fig. 6(a)

shows the test results of the DETR model and Fig. 6(b) shows the test results of the improved model. From Fig. 6(a), it can be seen that the original model will have missed detection as well as detection errors when there are more targets. The addition of CBAM attention and focus loss effectively improves the model's missed detection as well as the occurrence of detecting wrong targets, which proves the effectiveness of the improved model.

5. Summarize

In this paper, the Transformer model is applied to the weed detection task, based on the DETR model, channel attention and spatial attention are added to the DETR feature extraction network to improve the feature extraction ability of the model, and the focus loss function is used to improve the small target detection effect of the model. The final detection effect is improved by 1.12% based on the original DETR model. Comparing the before and after effects, our model improves the leakage detection of the DETR model and improves the ability of the model for small target detection.

References

[1] CHEN K, YANG H, WU D, et al. Weed biology and management in the multi-omics era: Progress and perspectives[J]. Plant Communications, 2024.

[2] ZHANG Z, LI R, ZHAO C, et al. Reduction in weed infestation through integrated depletion of the weed seed bank in a rice-wheat cropping system[J]. Agronomy for sustainable development, 2021, 41(1): 10.

[3] CHEN L, JIN M, ZHANG W L, et al. Research advances on characteristics, damage and control measures of weedy rice[J]. 2020.

[4] YUAN Hongbo, ZHAO Nudong, CHENG Man. Research progress and prospect of field weed recognition based on image processing[J]. Journal of Agricultural Machinery, 2020, 51(S2): 323-334.

[5] MATHANKER S K, WECKLER P R, TAYLOR R K, et al. AdaBoost and support vector machine classifiers for automatic weed control: Canola and Wheat[C]//2010 Pittsburgh, Pennsylvania, June 20-June 23, 2010. American Society of Agricultural and Biological Engineers, 2010: 1.

[6] Miao Ronghui, Yang Hua, Wu Jinlong, et al. Recognition of overlapping leaves and weeds in spinach based on image chunking and reconstruction[J]. Transactions of the Chinese Society of Agricultural Engineering, 2020, 36(4).

[7] B. Zhong, J. Yang, Y. Liu, et al. Weed detection in pasture based on improved DINO [Z]. Henan Agricultural Science, 2023, 52(9): 156-163.

[8] JIN X, LIU T, MCCULLOUGH P E, et al. Evaluation of convolutional neural networks for herbicide susceptibility-based weed detection in turf[J]. Frontiers in Plant Science, 2023, 14: 1096802.

[9] MOAZZAM S I, KHAN U S, QURESHI W S, et al. Towards automated weed detection through two-stage semantic segmentation of tobacco and weed pixels in aerial Imagery[J]. Smart Agricultural Technology, 2023, 4: 100142.

[10] LI J, ZHANG W, ZHOU H, et al. Weed detection in soybean fields using improved YOLOv7 and evaluating herbicide reduction efficacy[J]. Frontiers in Plant Science, 2024, 14: 1284338.

[11] ONG P, TEO K S, SIA C K. UAV-based weed detection in Chinese cabbage using deep learning[J]. Smart Agricultural Technology, 2023, 4: 100181.

[12] KHAN S D, BASALAMAH S, LBATH A. Weed–Crop segmentation in drone images with a novel encoder–decoder framework enhanced via attention modules[J]. Remote Sensing, 2023, 15(23): 5615.

[13] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 213-229.

[14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[15] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[16] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[17] JIANG H, ZHANG C, QIAO Y, et al. CNN feature based graph convolutional network for weed and crop recognition in smart farming[J]. Computers and electronics in agriculture, 2020, 174: 105450.