IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 24 July 2023, accepted 11 September 2023, date of publication 18 September 2023,
date of current version 26 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3316364

RESEARCH ARTICLE

Tobacco Leaf Segmentation Based on Improved MASK RCNN Algorithm and SAM Model

WEIZHENG ZHANG[1], YUEFENG WANG[1], GUANGCAI SHEN[2], CANLIN LI[1], MENG LI[3], AND YINGCHENG GUO[2]

[1]College of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou 450001, China
[2]Yunnan Tobacco Company, Baoshan Branch, Baoshan 678000, China
[3]College of Tobacco Science and Engineering, Zhengzhou University of Light Industry, Zhengzhou 450001, China

Corresponding author: Yingcheng Guo (bsycgyc@126.com)

This work was supported in part by the Henan Province Science and Technology Research Project 222102210037.

ABSTRACT High-precision segmentation of tobacco leaves is a prerequisite for analysis of phenotypic information. Challenges such as mutual occlusion and fuzzy edges make leaf segmentation difficult. This paper proposes an improved algorithm based on the Mask Region-based Convolutional Neural Networks (MASK RCNN) model and an instance segmentation method based on the SAM model to address these challenges. First, the MASK RCNN model is enhanced by incorporating a feature fusion layer and a hybrid attention mechanism, which improves the segmentation performance. The improved MASK RCNN model achieves an Avg.MIoU metric of approximately 85.10%, which is an improvement of 11.10% over the original algorithm. It also achieves an Avg.MPA metric of about 84.94%, indicating an improvement of 10.84%. Second, the Segment Anything Model (SAM) model is presented for the first time for tobacco leaf segmentation, providing empirical support for its application in the tobacco field. The SAM model demonstrates accurate segmentation of tobacco leaf images at different growth stages, demonstrating its good generality. In conclusion, the proposed methods effectively address the challenges in tobacco leaf segmentation, resulting in improved accuracy and performance. These techniques provide significant technical support for tobacco leaf phenotype research.

INDEX TERMS Tobacco leaf, occlusion, mask region-based convolutional neural networks (MASK RCNN), SAM, image segmentation.

I. INTRODUCTION

In plant phenotype research, obtaining accurate measurements of organ-associated phenotypes is critical [1]. Leaf segmentation is a key component in the acquisition of plant phenotype information [2]. Automating leaf instance segmentation and accurately extracting leaf shape has emerged as a prominent direction in plant phenotype research [3]. Traditional methods for leaf phenotype extraction rely on manual measurements and expert knowledge, which are time-consuming and subjective. With the advancement of plant functionalomics and breeding research, traditional phenotype observation has become a major bottleneck hindering progress. High-precision segmentation of plant stems and leaves is a challenging problem in plant phenotype research due to the blurring phenomenon. Plant phenotypic analysis is a scientific research field that involves the quantitative measurement of observable plant traits in response to the dynamic interaction between genotype and environmental conditions. It plays a critical role in understanding the effects of the environment on cultivated plants and has wide applications in areas such as plant breeding [4], crop monitoring [5], and disease prevention [6]. Traditional analysis of plant phenotypes relies on labor-intensive and error-prone manual measurements. However, the development of digital imaging and computer vision technologies allows non-intrusive and automated quantification of plant traits from images. Automatic segmentation of plant leaves is a fundamental requirement for achieving image-based plant phenotyping goals. Leaf segmentation can typically be performed at

The associate editor coordinating the review of this manuscript and approving it for publication was Qingli Li.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

two levels of granularity: category-level and instance-level. Category-level segmentation primarily involves separating pixels belonging to the leaf category from the background, while instance-level segmentation goes further by segmenting individual leaves from each other. Instance-level leaf segmentation enables fine-grained measurements of individual leaf area, leaf count, and leaf growth rate, which are highly beneficial for responsive plant growth monitoring and regulation. However, variability in leaf shape and appearance, persistent self-occlusion, and varying imaging conditions often pose significant challenges to instance-level leaf segmentation, even in controlled environments. In contrast, category-level leaf segmentation is relatively simpler and can well approximate the overall size of plants. Therefore, in many application scenarios, such as plant growth monitoring [7] and yield prediction [8], category-level leaf segmentation is a more feasible and practical approach. In summary, plant phenotypic analysis involves the quantitative measurement of plant traits that result from the interaction between genotype and environmental factors. The use of digital imaging and computer vision techniques enables the automated segmentation of plant leaves, providing valuable references for plant growth monitoring and yield prediction.

A. CONTRIBUTIONS

To address the above issues, this paper focuses on tobacco, a model plant, and proposes two efficient and accurate methods for tobacco leaf segmentation. The first method is an improved algorithm based on the MASK RCNN (Mask Region-based Convolutional Neural Networks) model. The second method introduces the Segment Anything model (SAM), which is applied to tobacco leaf segmentation for the first time. By exploring these two approaches, this study aims to improve the efficiency and accuracy of tobacco leaf segmentation and provide valuable insights for plant phenotype research.

The main contributions of this work are as follows: The improved algorithm, which is based on the MASK RCNN model, has the following advantages.

- A feature fusion layer introduced in the MASK RCNN model integrates original features, fractal features, and edge features.
- The omnidirectional gradient-based stylized edge extraction algorithm is used to extract the edge texture of tobacco leaves. With a wider perceptual field, it can capture edge relationships between distant pixels.
- Fractal features have the ability to represent multiple scales, model self-similarity, and are not limited by image resolution. They also exhibit robustness to noise and disturbances in the data.
- In the feature encoding stage of the mask segmentation network structure of MASK RCNN, a hybrid attention mechanism is added to effectively combine the channel attention mechanism with the spatial attention mechanism.

The tobacco segmentation algorithm based on the SAM model has the following advantages.

- Powerful Perceptual Field, the SAM model is able to capture the spatial dependencies between different locations in an image within the perceptual field through a self-attentive mechanism.
- The SAM model exhibits a flexible feature representation similar to Transformer. It can dynamically learn the importance of different locations in an image and focus on the correlation between various elements, thereby adjusting the feature representation adaptively.
- The SAM model employs a multi-scale feature fusion strategy in the decoder section. This approach integrates features from various levels of the encoder to generate a more comprehensive and multi-scale feature representation.
- The SAM model is designed with a focus on computational efficiency, taking into account the avoidance of complex operations commonly found in traditional methods.

The remaining part of the paper is organized as follows:

Section II presents the literature survey of related works. Section III presents the basic framework and improvement process of MASK RCNN is introduced Section IV presents the SAM model to realize the process of tobacco segmentation. The analysis of the segmentation effect and accuracy of two models is presented in Section V. Subsequently, the improved algorithm is discussed in section IV.

II. RELATED WORK

In recent years, both domestic and international scholars have conducted extensive research on leaf instance segmentation. Traditional methods in this field mainly rely on classical image processing techniques. For example, Pape et al. [9] proposed a method that combines leaf color and texture features for leaf segmentation, Viaud et al. [10] utilized the watershed method for leaf instance segmentation, and Yin et al. [11] extended the multi-leaf alignment and tracking framework to instance segmentation using chamfer matching. However, these traditional methods often have limitations in terms of accuracy.

With the advancements in deep learning, researchers have started applying deep learning models to leaf instance segmentation to improve accuracy and performance. The Leaf Segmentation Challenge held at the European Conference on Computer Vision in 2014 provided a publicly available dataset (CVPPP) for related studies in plant phenotyping [6]. Scholars have conducted numerous studies on leaf instance segmentation strategies using this dataset. Romera et al. [12] initially proposed a recurrent instance segmentation algorithm based on recurrent neural networks but achieved limited results. Subsequently, some researchers improved this algorithm by incorporating Conditional Random Fields (CRF) as post-processing, which led to improved segmentation results, yet the overall segmentation effect remained restricted. Ren et al. [13] introduced an end-to-end

recurrent neural network architecture with an attention mechanism specifically designed for leaf instance segmentation. This algorithm successfully addressed partial occlusion problems and significantly improved fine segmentation outcomes. Another study by Ward et al. [7] explored the use of synthetic data for plant imaging. They synthesized 3D plant models through domain randomization, generated rendered synthetic data samples, and employed the Mask RCNN model for training on the synthetic dataset. Kuznichov et al. [14] proposed a copy-and-paste data augmentation algorithm based on the Mask RCNN model. This technique involves copying selected regions of interest from source images and pasting them into random locations within the same image while blending them with the target location based on mask information. This operation facilitates the generation of more training samples, leading to improved performance of the Mask RCNN model for plant detection and segmentation tasks in agricultural settings. However, it is worth noting that these methods often require extensive image preprocessing as part of the image enhancement techniques involved.

Overall, the application of deep learning models has shown promising results in improving the accuracy of leaf segmentation. While some approaches necessitate image preprocessing, they have demonstrated remarkable advancements in this field. These studies have made significant progress in image segmentation using MASK RCNN models, but they have also encountered some limitations. For example, MASK RCNN models are highly dependent on annotated data and require a large amount of training data with accurate segmentation annotations to achieve good performance. However, acquiring and annotating such datasets can require a lot of work and expertise. The results are poor for small targets. Because the MASK RCNN model uses the Region Proposal Network, the network may not be accurate enough to localize and segment small targets. This is because it is often difficult to generate stable candidate frames for small targets. Unstable processing of complex backgrounds: The MASK RCNN model may miss segmentation or segmentation when processing images with complex backgrounds. This is because texture and color variations in complex backgrounds can interfere with the model's ability to accurately detect targets.

This paper addresses the existing limitations of MASK RCNN at the current stage and proposes corresponding improvements. MASK RCNN is a deep learning algorithm specifically designed for instance segmentation, which can simultaneously perform multi-target detection and instance segmentation. In this paper, based on MASK RCNN, a boundary extraction algorithm is proposed to generate effective boundary features to improve the clarity of the tobacco leaf boundary. In addition, the fractal features are fused into the network by using the fractal dimension to represent the self-similarity of tobacco leaves at different scales, which improves the fractal feature description capability of the network. A feature fusion layer is also added to the network to fuse the boundary features and fractal features into the original feature map. Meanwhile, the FCN (Fully Convolutional Network) structure in MASK RCNN is adjusted and a hybrid attention module is introduced to reduce the loss of detail information of the tobacco leaves caused by the convolutional operation. In addition, this paper also applies the SAM to tobacco segmentation for the first time and demonstrates that it can be successfully applied to the agricultural field, i.e., tobacco segmentation.

III. TOBACCO INSTANCE SEGMENTATION BASED ON IMPROVED MASK RCNN MODEL

A. MASK RCNN MODEL BASE FRAME

The backbone network of the MASK RCNN algorithm [15] usually uses a residual network, ResNet, to gradually extract the low-level features and high-level features of the leaf images from the bottom layers to the top layers. Then, the feature pyramid networks (FPNs) are used to transfer the high-level features to the bottom-level features for feature fusion and to form the feature map into the region proposal networks (RPNs). The RPN searches the feature map for regions containing the target at different scales and generates region proposal boxes. For each proposal box, the RPN outputs two results, a foreground/background classification and a foreground bounding box. Next, the suggestion boxes are classified and a mask and bounding box are generated. The classifier is used for specific classification of the suggestion frame and fine tuning of the bounding box to achieve the target detection function. Finally, a Fully Convolutional Network (FCN) is used to generate the mask for the suggestion frame, complete the image instance segmentation, and generate a pixel-level mask for the target. The above is the main process of the MASK RCNN algorithm, which achieves accurate target detection and instance segmentation through multi-stage processing combined with feature fusion and region suggestion network. The specific network structure is shown in Fig.1.

B. ENHANCED MASK RCNN

1) EXTRACTION OF BOUNDARY FEATURES

There is a covering phenomenon between the tobacco leaves, and the leaves have low contrast, which leads to unclear boundaries between the leaves, this paper refers to the stylized edge extraction algorithm based on omnidirectional gradient [16], the method based on omnidirectional gradient can efficiently extract the edge information of the image and can be combined with the visual effect of human perception of the stylized processing. The algorithm is computed by using a flexible convolution kernel radius and a special law, as long as the convolution radius r is large enough, the gradient direction of a pixel is accurate enough, and the omnidirectional gradient can be obtained by synthesizing the gradient values in multiple directions. And it overcomes the disadvantage of classical edge extraction algorithms that require manual thresholding. Classical edge detection operators such as Soble and Canny can only detect edges in two directions:

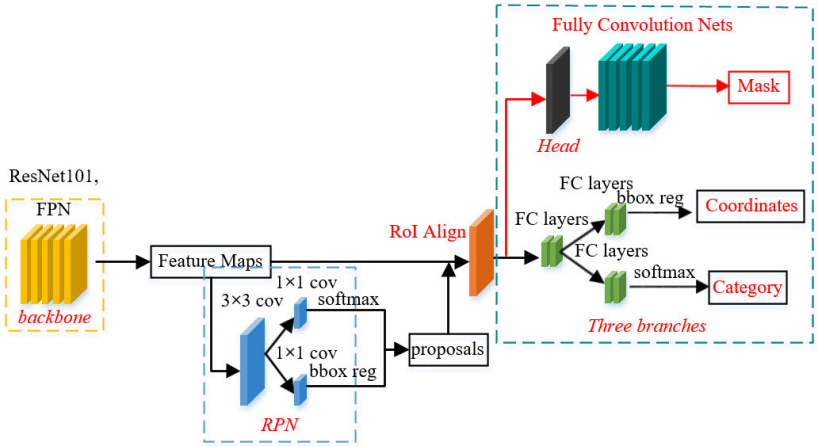


FIGURE 1. MASK RCNN structure diagram.

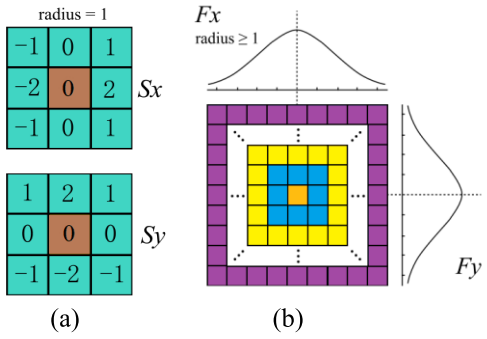


FIGURE 2. Comparison of Convolution Factors. (a) Soble edge detection operator (b) Method in this paper.

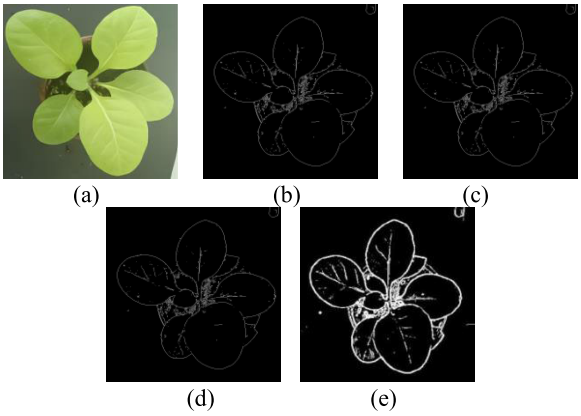


FIGURE 3. Comparison of boundary extraction algorithms.(a) original picture (b) Prewitt (c) Roberts (d) soble (e) this paper.

horizontal and vertical. The example of the convolution factor pair is shown in Fig.2.

As shown in equation (1), G_θ can be normalized to 8 gradient directions.

$$G_\theta = \begin{cases} arc\left(G_y^0/G_x^0\right) & G_x^0 > 0 \\ arc\left(G_y^0/G_x^0\right) + \pi & G_x^0 < 0, G_y^0 \geq 0 \\ arc\left(G_y^0/G_x^0\right) - \pi & G_x^0 < 0, G_y^0 < 0 \\ 0 & G_x^0 = 0, G_y^0 = 0 \\ +\pi/2 & G_x^0 = 0, G_y^0 > 0 \\ -\pi/2 & G_x^0 = 0, G_y^0 < 0 \end{cases} \quad (1)$$

The extraction results of the classical edge-detection operator with the method of this paper are shown in Fig.3.

2) FRACTAL FEATURE EXTRACTION

To further improve the segmentation accuracy of the MASK RCNN network for tobacco, the quality of feature extraction is particularly important. If the features of the region to be segmented are not obvious, the segmentation accuracy of the network will be reduced. Therefore, this paper introduces fractal features into MASK RCNN to improve the segmentation accuracy. In order to distinguish different leaves, the size of the fractal dimension and the distribution in a certain region can be used for calculation. In this paper, we use the fractal interpolation function to map the image into a rational fractal surface and calculate its box dimension to obtain the fractal dimension [17]. The specific calculation method is shown below.

$$S = \begin{Bmatrix} \left|s_\tau{}^{-1}(1)\right| & \left|s_\tau{}^{-1}(1)\right| & \cdots & \left|s_\tau{}^{-1}(1)\right| \\ \left|s_\tau{}^{-1}(2)\right| & \left|s_\tau{}^{-1}(2)\right| & \cdots & \left|s_\tau{}^{-1}(2)\right| \\ \vdots & \vdots & \vdots & \vdots \\ \left|s_\tau{}^{-1}(N^2)\right| & \left|s_\tau{}^{-1}(N^2)\right| & \cdots & \left|s_\tau{}^{-1}(N^2)\right| \end{Bmatrix}_{N^2 \times N^2} \quad (2)$$

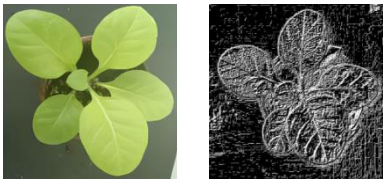
As shown in equation (2), S is the scale factor matrix; $\tau(i,j)=(i-1)\times N+j$ representing the enumeration of set $\{(i,j) : i,j = 1,2,\ldots,N\}$, and $\tau^{-1}(N)$ is used to map $\tau(i,j)$ to position (i,j). The fractal dimension's box dimension D is denoted:

$$D = \begin{cases} 1 + \log_N \lambda, & \lambda > N, \\ 2, & \lambda \leq N. \end{cases} \quad (3)$$

$$\lambda = \rho(S) = \sum_{k=1}^{N^2} s_{\tau(k)}^{-1} = \sum_{i=1}^{N}\sum_{j=1}^{N} s_{i,j} \quad (4)$$

$S_{i,j}$ as a scale factor, then equation (3) can be transformed to be:

$$D = \begin{cases} 1 + \log_N \sum_{i=1}^{N} \sum_{j=1}^{N} |s_{i,j}|, & \sum_{i=1}^{N} \sum_{j=1}^{N} |s_{i,j}| > N, \\ 2, & others. \end{cases} \quad (5)$$



(a) (b)

FIGURE 4. Fractal feature extraction map. (a) original picture (b) feature image.

The fractal dimension of the entire image is mapped onto a matrix of fractal features, and then the image of the fractal features is displayed, as shown in Fig.4.

3) FEATURE FUSION

To fuse boundary features and fractal features into the feature map, a feature fusion layer is introduced. This layer consists of two parts, the cropping layer and the fusion operation, which are used to fuse the boundary features, fractal features, and original features. Specifically, we first crop the boundary feature map and the fractal feature map using a 1×1 convolution operation to ensure that their feature dimensions are the same as the original feature map. Next, the dimension sizes between different feature maps are unified by operations such as up-sampling and down-sampling. Then, the trimmed boundary feature maps, fractal feature maps, and original feature maps are fused. As shown in Fig.5.

In this paper, we use the add feature fusion operation to realize the feature fusion. Specifically, we fuse the boundary featuresand the new fractal features into the original feature map by the add operation, and the fusion formula (6) is as follows:

$$F_{new} = F_{boundary} \oplus D_{fractal} \oplus F_{original} \quad (6)$$

where $F_{boundary}$ denotes boundary features, $D_{fractal}$ denotes fractal features, and $F_{original}$ denotes model base features. This feature fusion approach can make full use of different levels of feature information, enrich the underlying features, increase the feature information, and enhance the detailed features while preserving the background information, thus improving the performance of the model.

4) MASK RCNN SEGMENTATION NETWORK ARCHITECTURE

In the MASK RCNN mask segmentation network, the feature map obtained from the ROIAlign layer is first fixed to a size of $14 \times 14 \times 256$ by a pooling layer, and then four convolution operations are performed on it. Next, an inverse convolution operation is performed on the fifth layer to obtain the mask, and the number of channels in the mask is adjusted by convolution to match the number of target species. Although the MASK RCNN can recover the category to which the pixel belongs, it is poor at recognizing the edges of the tobacco leaf. This is due to the fact that after multiple convolution operations, the resolution of the image is reduced and the detail information is lost. Although the resolution of the image can be gradually recovered in FCN networks using four deconvolution operations, the single deconvolution operation can only recover the feature layer to a certain extent, while it cannot recover some of the lost image information. In addition, the four-layer simple deconvolution operation will further increase the error, resulting in inaccurate segmentation of the original tobacco leaf. By adding a hybrid attention module to the segmentation network, the multi-scale feature information can be better utilized and the semantic information of the image can be fully explored, thus obtaining a more accurate tobacco leaf segmentation mask. The MASK RCNN segmentation network structure is shown in Fig.6.

5) HYBRID ATTENTION MECHANISM

The channel attention mechanism [18] mainly adjusts the weight of each channel by modeling the relationship between different channels of the feature map, so as to utilize the channel information more effectively. The spatial attention mechanism [19], on the other hand, adjusts the weight of each region by modeling the relationship between different regions of the feature map in order to exploit the spatial information more effectively. The combination of the channel attention mechanism and the spatial attention mechanism allows simultaneous attention in both the channel and spatial dimensions of the feature map.

The hybrid attention module is shown in Fig.7. There are four parts in this module, and this module is mainly composed of four parts, which are obtaining channel attention coefficients, obtaining spatial attention coefficients, weighted fusion of attention coefficients, and multiplication of attention coefficients with the original feature layer. First, the channel attention coefficients and spatial attention coefficients are obtained by the channel attention mechanism and the spatial attention mechanism, respectively. Then, these two attention coefficients are weighted and merged to obtain the final attention coefficients. Finally, the attention coefficients are multiplied by the original feature layer element by element to obtain the feature map with enhanced channel and spatial information representation.

After a comprehensive consideration of model performance, computational efficiency, dataset characteristics, model structure and architecture, and prior knowledge, this paper adopts a hybrid attention mechanism.The combination of channel and spatial attention mechanisms provides more comprehensive and flexible attention capabilities, reduces the interference of redundant information, improves the robustness and generalisation of the model, improves the computational efficiency of the network, and has wide applicability.

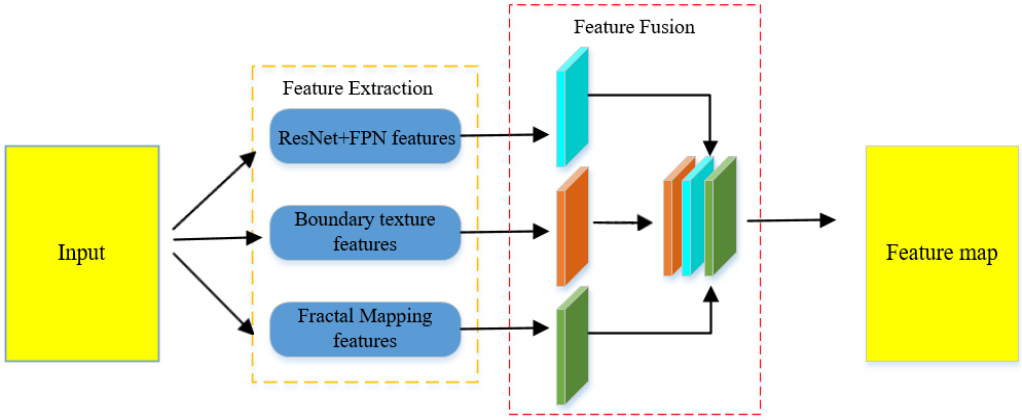


FIGURE 5. Feature fusion.

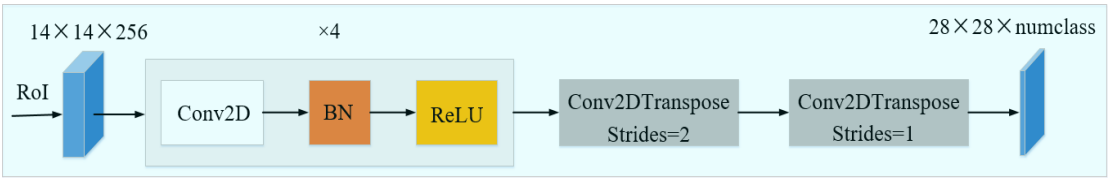


FIGURE 6. MSAK RCNN mask segmentation network structure.

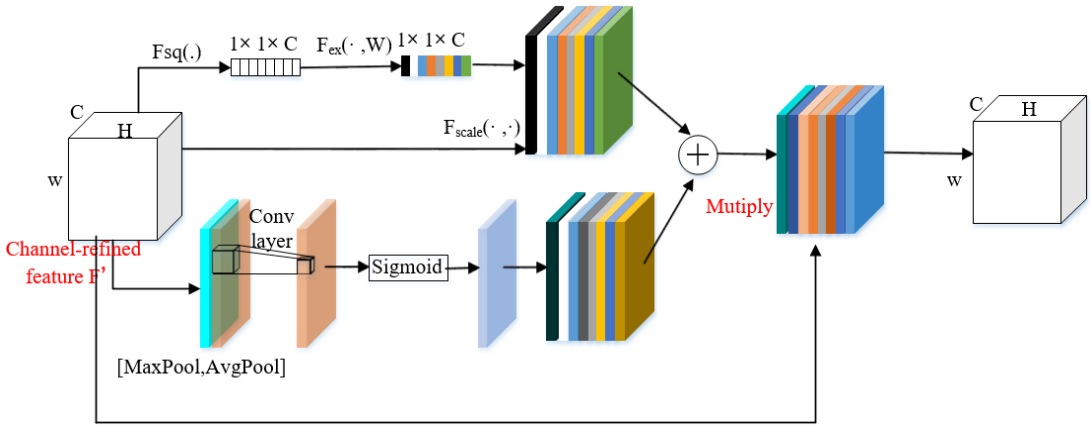


FIGURE 7. Hybrid attention module.

This gives it a unique advantage among current state-of-the-art attention modules.

The hybrid attention module is added to the feature encoding stage of the segmentation network to better enhance the semantic information of the tobacco leaf and gradually recover an accurate tobacco leaf segmentation mask using multi-scale feature layers.

IV. SEGMENTATION OF TOBACCO INSTANCES BASED ON THE SAM MODEL

A. DIFFERENCE BETWEEN SAM MODEL AND TRADITIONAL MODEL DATA ANNOTATIONS

The SAM [20] is a self-supervised learning-based image segmentation method, which is significantly different from the traditional model data labeling methods in terms of whether or not manual data labeling is required. Traditional model data labeling usually requires a large amount of labor and time, and requires professional labelers to manually label each sample. Although this method can achieve more accurate labeling results, it also has some problems, such as unstable labeling quality, slow dataset update, and high labeling cost. In contrast, the SAM model adopts a self-supervised learning approach to train the model using existing unlabeled data. This approach allows the model to automatically learn both the feature representation and the target segmentation masks from the data itself. Specifically, the SAM model introduces perturbations to the input image through techniques like random cropping and rotation, creating multiple deformed versions of the image. These deformed versions are then used as input to predict the corresponding segmentation

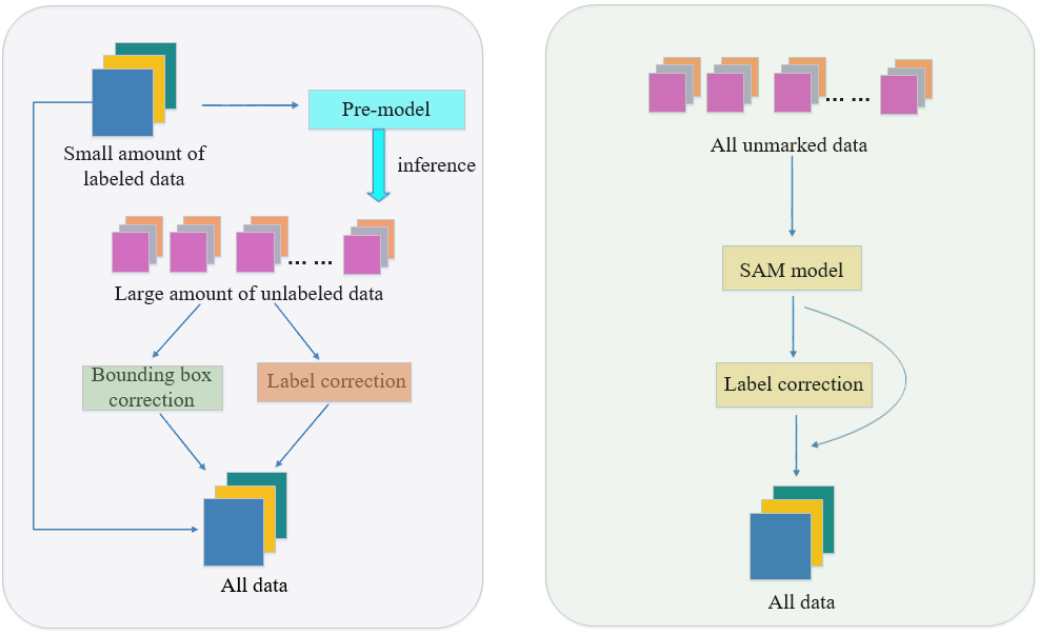


FIGURE 8. Generic automated annotation methods are shown left and SAM model-based annotation methods shown right.

masks using unsupervised learning techniques. By aggregating the predictions from these multiple versions, more accurate segmentation results can be obtained. Compared to traditional data labeling methods, the SAM model offers several advantages. Firstly, it does not require manual annotation of large amounts of data, saving labor and time. Secondly, it leverages existing unlabeled data, improving data utilization and efficiency. This approach enhances the scalability of the model and reduces the dependence on labeled training data.

Fig.8 illustrates the workflow of both traditional data labeling methods and the self-supervised learning approach employed by the SAM model, highlighting the differences in their data processing pipelines.

B. SAM MODEL ARCHITECTURE

The SAM model contains three main components: a powerful image encoder (used to compute image embeddings), a cue encoder (used to compute cue embeddings), and a lightweight mask decoder (used to predict masks in real time). These components work together to form the overall architecture of the SAM model.

- Image encoder: The researchers used the pre-trained MAE (Mixup AutoEncoder) ViT (Vision Transformer) as the image encoder, which is based on scalable and robust pre-training methods that are minimally suited for processing high-resolution inputs. The image encoder is run once for each image and applied before the cueing model.
- Cue Encoder: The SAM model considers two types of cues: sparse (dots, boxes, text) and dense (masks). For sparse cues, the researcher represents dots and boxes using positional encoding, and sums the learned embeddings and free-form text using standard text encoding in CLIP. For dense cues (i.e., masks), embeddings are performed using convolution and elements are summed using image embedding.
- Mask Decoder: The mask decoder of the SAM model effectively maps image embeddings, cue embeddings, and output tokens to masks, allowing real-time prediction of masks. The mask decoder plays a key role in the entire SAM model to produce high-quality segmentation results by effectively fusing image and cue information.

C. SAM MODEL SEGMENTATION IMAGE FLOW

SAM is a versatile image segmentation base model that supports multiple input cues to improve segmentation quality. To facilitate comparison and evaluation, SAM uses the center of mass of the ground truth mask for each instance as a cue for segmenting each instance. After receiving these cues, SAM generates three potential segmentation results and provides a corresponding score for each result. Then, the result with the highest score is selected and evaluated by comparing it to the ground truth mask. Algorithm 1 describes the details of the implementation.

SAM is a suggestive model that has been trained on over 11 million images and has generated 1 billion masks. This large-scale training allows the SAM model to have strong zero-sample generalization capabilities, i.e., accurate image segmentation can still be performed for unseen targets or scenes. It can be used for a variety of

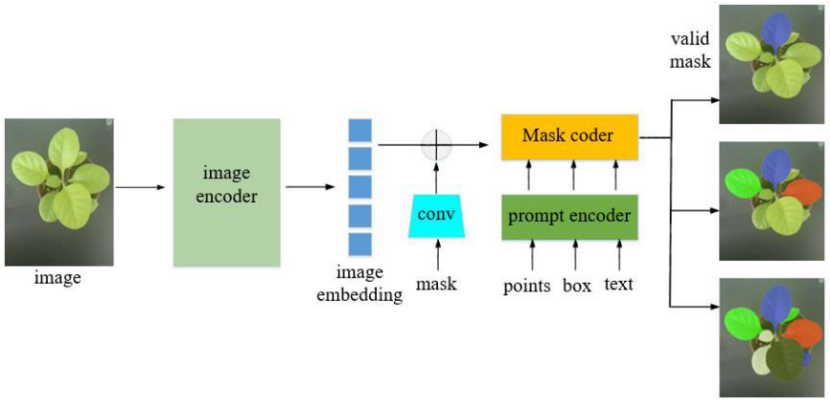


FIGURE 9. SAM model.

Algorithm 1 Application of the SAM Model to Tobacco Segmentation

Input: *a pre-trained SAM model denoted as A(·, ·), a contour detector denoted as CD(·), a midpoint detector denoted as MD(·), a tobacco image dataset denoted as I containing category labels C, each image in the dataset denoted as i, and each category cls in the set C of category labels:*
Output: *the set M of segmented masks.*
1: ***for*** *i ∈* ***Ido***
2: ***for*** *cls ∈* ***Cdo***
3: $i_c \leftarrow$*(label(i) == cls)*
4: Contours$\leftarrow$ *CD* (i_c)
Initialize image mask m
5: for c∈ Contours do
6: P∈ MD(c)
7: mouts, scores$\leftarrow$ *A (i, P)*
8: m$\leftarrow$ *Argmax(scores(mouts))*
9: ***end***
10: M.append(m)
11: ***end***
12: ***end***

image segmentation tasks, including semantic segmentation, instance segmentation, contour detection, etc. Fig.9 shows the flowchart of the SAM model for image segmentation.

V. EXPERIMENTS AND RESULTS ANALYSIS

A. DATASET AND ENVIRONMENT

In this paper, we utilized the CVPPP plant leaf segmentation dataset [6], which comprises five distinct sub-datasets (A1, A2, A3, A4, and A5). The A5 dataset is a combination of A1 to A4 datasets. For our experimentation purposes, we selected 300 images from the A5 dataset. Moreover, we supplemented the dataset with 300 images of tobacco leaves collected at different growth stages. These tobacco leaves were grown under controlled environmental conditions with temperature settings ranging from 15 to 20 degrees Celsius, along with supplementary lighting. For image acquisition, we employed standard high-resolution digital cameras equipped with zoom lenses (specifically, the Canon Cannon 600D model) and diffuse fill lights. To enhance image clarity and minimize shadows, two diffuse fill lights were used in conjunction with a black background. These measures resulted in sharper images devoid of shadows.

B. EXPANDED DATA SET

The main challenge in utilizing convolutional neural networks, such as for segmentation tasks, lies in collecting and labeling a sufficient number of training samples. In our case, the combined dataset of self-collected images and publicly available images comprises a total of 500 images, which may be insufficient to meet the requirements for effective deep learning models and can lead to overfitting. To address this challenge, data augmentation techniques were applied to enhance the original dataset. This involved randomly mirroring, flipping, and resizing the images, resulting in an expansion of the training set from 500 to 2000 images. Additionally, the number of self-collected images was expanded to 1200, and the image sizes were standardized to ensure consistency in the input image dimensions. Considering the constraints of graphics card memory, the images in the training and test datasets were uniformly resized to 800 $\times$ 800 pixels.

C. BUILDING A TRAINING PLATFORM

The models in this study were trained and tested on a computer equipped with an 8-core CPU running at 3.50 GHz, 32 GB of RAM, Nvidia GTX2080Ti GPU, and Ubuntu 16.04 operating system accelerated with a GTX2080Ti GPU. The leaf segmentation model was deployed based on the open source mmdetection and pytorch frameworks, configured to install a Python 3.7 environment, Cuda 11.0 computational architecture, and Cudnn 7.6 acceleration library.

The Mask RCNN network is trained with the network parameters shown in Table 1. First, the images of the training set are resized to 600 $\times$ 600px, and then the network parameters of the pre-trained model are set. The RPN anchor is set to (16, 32, 64, 128, 256), the number of iterations (max epoch)

TABLE 1. Network parameters.

Parameter name	Set Parameters
Initial learning rate	0.01
momentum factor	0.9
training rounds	200
optimizer	Adam
activation function	ReLU
Output image size/px	800×800

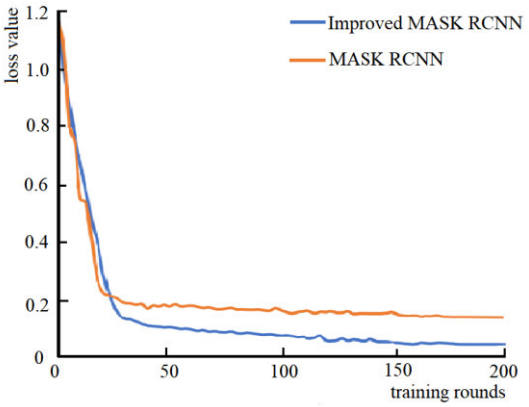


FIGURE 10. Model training loss value plot.

is set to 200, the initial learning rate is set to 0.01, and the momentum factor is set to 0.9.

The variation curve of the loss function value of the improved mask RCNN network is shown in Fig.10. The horizontal coordinate indicates the number of training iterations, and the loss value of the model has been decreasing as the number of iterations increases. The loss value decreases faster in the first 10 iterations, indicating that there is a large gap between the parameters of the model and the optimal parameters. The parameters of the model are continuously optimized to approximate the optimal values by iterating the training set. The number of iterations between 10 and 150, the loss value decreases more slowly, indicating that the model parameters are close to the optimal value at that time. To avoid the overfitting phenomenon, the model parameters can be fine-tuned by reducing the learning rate. When the number of iterations reaches 150, the loss value of the training set stabilizes at about 0.025, indicating that the model training has basically converged at this time and the training effect is good.

D. TWO MODELS RESULTS AND ANALYSIS

Fig.11 shows the segmentation effect graph of the original MASK RCNN. Typical segmentation types selected from the test dataset are shown, with the presence of unclear leaf edges, poor segmentation of the petiole portion (petiole is too thin), and small leaves not segmented in the case of

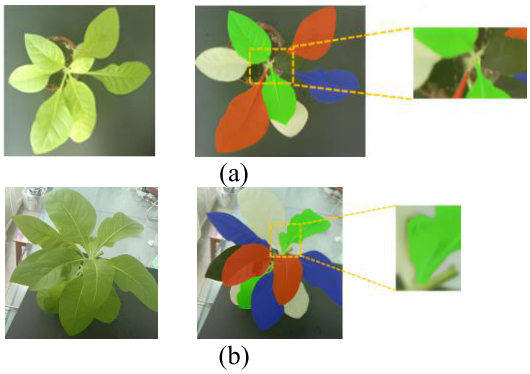


FIGURE 11. Original MASK RCNN segmentation effect.

mutual occlusion. The original images for the two cases correspond to figures (d) and (e) in Fig.12, respectively.

Fig.12 illustrates the morphology of tobacco leaves at different growth stages, namely seedling, vegetative, and lush. Each stage presents distinct morphological characteristics: Seedling stage leaves are small, thin and soft, and are usually heart-shaped or ovate, with possibly serrated edges. During the vegetative stage, the leaf becomes larger and broader, elliptical to lanceolate, relatively narrow in width, and may have a wavy edge. Lush Growth Stage: Tobacco leaves are essentially oval or obovate in shape, relatively wide, with distinct midrib and secondary veins and a thicker leaf texture. In the original Fig.12(a), some of the leaves are small and one of them is heavily shaded. In contrast, in the original Fig.12(b), the petiole is slender and the unimproved MASK RCNN model does not accurately segment the petiole. In addition, in the original Fig.12(e), some of the leaves are not fully revealed due to the shooting angle and the occlusion between the leaves, and the unimproved MASK RCNN model does not fully segment these leaves, and the segmented edge effect is obviously not clear.

Fig.13 and Fig.14 show the performance of the improved MASK RCNN model and the SAM model in segmenting tobacco leaves at each growth stage. First, the tobacco leaf image is segmented using the enhanced MASK RCNN model, followed by segmentation using the SAM model. The enhanced MASK RCNN model incorporates a feature fusion layer to combine extracted image stylized edge features, fractal features, and original features. In addition, a hybrid attention mechanism is introduced to improve image segmentation. This enhancement enables the model to effectively capture information across different scales and dimensions, thereby improving segmentation accuracy and robustness. On the other hand, the SAM image segmentation model improves segmentation accuracy and robustness by implementing technical means such as a spatial attention mechanism and an aligned feature representation. These techniques allow the model to focus on relevant regions and better represent feature patterns, resulting in improved segmentation performance. By introducing these enhancements, both the improved MASK RCNN model and the SAM model

FIGURE 12. Original image.

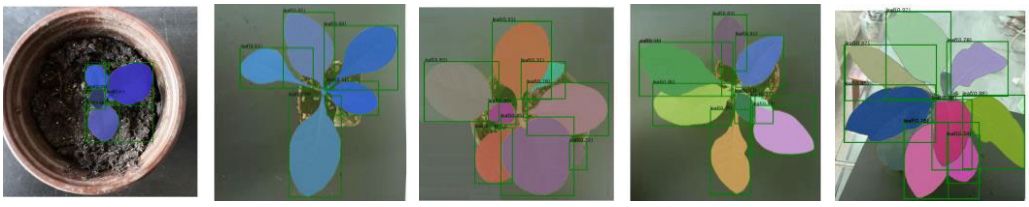


FIGURE 13. Segmentation effect of tobacco leaves by improved MASK RCNN.

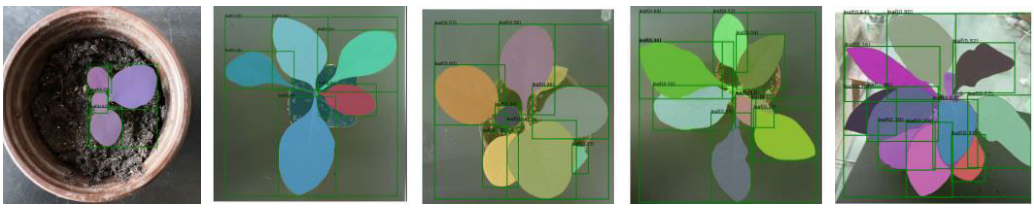


FIGURE 14. Segmentation effect of tobacco leaves by SAM.

demonstrate increased accuracy and robustness in tobacco leaf image segmentation.

By using the improved MASK RCNN model and the SAM model for tobacco image segmentation, whether for tobacco leaves of different growth periods, or for problems such as severe inter-leaf occlusion, difficult to capture edge information, thin and narrow petioles, or poorly selected shooting angles, each of the two models shows good segmentation results and successfully solves the problems in the above typical cases. Meanwhile, these segmentation results of tobacco images provide empirical support for the application of the SAM model in the tobacco field, verify the application potential of the SAM model in the tobacco field, and provide a basis for further research and development.

E. SEGMENTATION EFFECTIVENESS EVALUATION OF SAM AND IMPROVED MASK RCNN MODELS

1) EVALUATION INDICATORS

In segmentation algorithms, the following metrics are commonly used to evaluate segmentation effectiveness: Intersection and Fusion Ratio (IoU), Mean Intersection over Union (MIoU), Pixel Accuracy (PA), Category Pixel Accuracy (CPA), and Mean Category Pixel Accuracy (MPA).

These metrics can help evaluate the performance of segmentation algorithms in different scenarios, and they play an important role in guiding the training and tuning of segmentation models.

IOU: The accuracy of segmentation is evaluated by calculating the ratio of the intersection area to the merge area between the predicted and true values. The value of this ratio ranges from 0 to 1. The closer the value is to 1, the more similar the predicted and true results are, and the better the segmentation effect is. Equation (7) is as follows.

$$IoU = \frac{TP}{TP + FP + FN} \tag{7}$$

MIoU (Mean Intersection and Merger Ratio) is used to evaluate the average of the IOU values of the same category in the image prediction results and is calculated by the formula (8) as follows.

$$MIoU = \frac{1}{class} \sum_{i=1}^{class} IoU_i \tag{8}$$

PA is used to evaluate the number of correctly predicted pixels in the image prediction results as a percentage of the total image pixels, and is calculated as shown

in equation (9) below.

$$PA = \frac{TP + TN}{TP + FP + FN + TN} \tag{9}$$

CPA is used to evaluate the percentage of correct predictions for a single category in the image prediction results, i.e., the ratio of the number of correctly predicted image pixels to the sum of correctly predicted image pixels of that category and incorrectly predicted image pixels not of that category. The formula is shown below.

$$CPA = \frac{TP}{TP + FP} \tag{10}$$

MPA is the average percentage of pixels correctly predicted across all categories.

$$MPA = \frac{1}{class} \sum_{i=1}^{class} CPA_i \tag{11}$$

In this experiment, we mainly used MIoU and MPA as the evaluation indices of the network segmentation effect. Since different confidence threshold settings can lead to different prediction results, we set the confidence threshold to 0.50, 0.75, and 0.85 for three cases for experimental comparison. The corresponding evaluation metrics are MIoU50, MIoU75, MIoU85, MPA50, MPA75, and MPA85.In addition, we calculated the average values of MIoU and MPA under these three thresholds as the comprehensive evaluation metrics, which are denoted as Avg. MIoU and Avg.MPA. By using these evaluation metrics, we can comprehensively evaluate the segmentation effect of the model under different confidence thresholds, and compare and analyze the segmentation effects.

TABLE 2. Test results of five networks on MIoU50, MIoU75, MIoU85.

Network model	$MIoU_{50}$	$MIoU_{75}$	$MIoU_{85}$	Avg.MIoU
MASK RCNN	0.7521	0.7439	0.7242	0.7400
U-Net	0.7954	0.8052	0.8027	0.8011
Deeplab v3+	0.8120	0.8236	0.8132	0.8163
U-Net++	0.8279	0.8353	0.8328	0.8220
Attention U-Net	0.8384	0.8230	0.8223	0.8279
Improved MASK RCNN	0.8501	0.8653	0.8364	0.8510
SAM	0.8394	0.8532	0.8217	0.8381

2) COMPARISON BETWEEN FIVE NETWORKS

To demonstrate the superiority of our proposed improved MASK RCNN segmentation network and SAM segmentation network, we conduct experiments under the same environment. Specifically, we compare these models with several popular segmentation algorithms, including MASK RCNN, U-Net, DeepLabv3+, U-Net++, Attention U-Net. The performance of each algorithm is evaluated and compared based on their respective results. The detailed comparison results are presented in Table 2, while Fig.15 illustrates the bar graphs showcasing the MIoU test results for the seven networks. This visual representation allows for a clear visualization of the performance differences between the different algorithms. By conducting this comprehensive evaluation and comparing the performance of these segmentation algorithms, we aim to provide evidence supporting the superior performance of our proposed improved MASK RCNN segmentation network and SAM model.

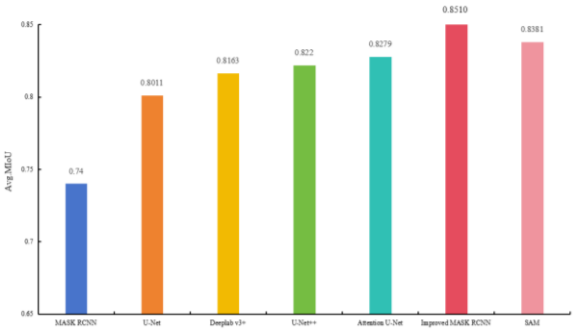


FIGURE 15. MIoU test results of five networks.

According to the results in Fig.15, using Avg.MIoU as the evaluation index of the segmentation network, the effectiveness of each algorithm is ranked as follows: MASK RCNN < U-Net < DeepLab v3+ < U-Net++ < Attention U-Net < SAM< Improved MASK RCNN. It can be seen that the MASK RCNN model has the worst segmentation effect; compared to the MASK RCNN model, the Improved MASK RCNN model improves about 11.10% on Avg.MIoU; compared to the MASK RCNN model, the SAM model improves about 9.81% on Avg.MIoU; the Improved MASK RCNN model is slightly higher than the SAM model by 1.29%. It can be seen that our proposed improved MASK RCNN model has a significant improvement in the segmentation effect, and the SAM model performs well in the segmentation effect.

TABLE 3. Test results of the five networks at MPA50, MPA75, MPA85 and MPA.

Network model	MPA_{50}	MPA_{75}	MPA_{85}	Avg.MPA
MASK RCNN	0.7412	0.7456	0.7363	0.7410
U-Net	0.8235	0.8128	0.7932	0.8098
Deeplab v3+	0.7936	0.8152	0.7984	0.8024
U-Net++	0.8226	0.8237	0.8179	0.8214
Attention U-Net	0.8301	0.8243	0.8407	0.8317
Improved MASK RCNN	0.8416	0.8527	0.8586	0.8494
SAM	0.8376	0.8421	0.8318	0.8371

Comparing the MPA test results of the four networks according to Table 3 and Fig.16, the following can be found. Figure 16 shows the bar chart of the MPA test results for the five networks. Among these four network structures, the segmentation effect of the unimproved algorithm is the worst, and its Avg.MPA is only 74.10%. While the other four networks have MPAs above 80%, the improved MASK RCNN has the highest Avg.MPA value of about 84.94%.

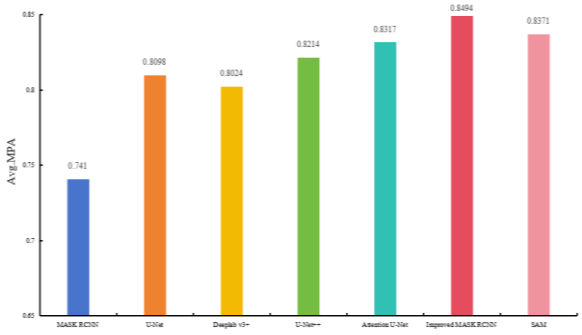


FIGURE 16. MIoU test results of five networks.

The MPA of the improved MASK RCNN is improved by about 10.84% compared to the MASK RCNN, the SAM model is higher than the MASK RCNN by about 9.61%, and the SAM model is slightly lower than the improved MASK RCNN by about 1.2%. It can be seen that the improved segmentation algorithm achieves significant improvement in MPA and performs best among all the networks. This indicates that our proposed improved segmentation algorithm has an advantage in terms of classification accuracy at the pixel level and can perform target segmentation more accurately.

The enhanced MASK RCNN model, trained on the designated leaf dataset, exhibits high confidence in target detection. It proves to be well-suited for both leaf detection and instance segmentation tasks, achieving accurate segmentation results. It takes approximately 30 seconds to complete the segmentation of a single image from the test set. On the other hand, the SAM model demonstrates lower confidence in tobacco leaf detection but still performs accurate image segmentation. Notably, it excels in handling the segmentation of unseen targets or scenes effectively. However, it typically requires around three minutes to perform multi-target instance segmentation. These findings highlight the trade-off between target detection confidence and segmentation accuracy in the two models. The improved MASK RCNN model provides faster inference times and higher target detection confidence, making it suitable for various applications. Meanwhile, the SAM model showcases its effectiveness in accurately segmenting images, particularly for novel or unseen targets, despite longer processing times.

VI. CONCLUSION

In this study, the challenges in leaf segmentation are thoroughly investigated and the existing model is improved. The improved segmentation algorithm achieved significant advantages compared with the original algorithm and other common algorithms, and the improved MASK RCNN model achieved a result of about 85.10% in the Avg.MIoU metric, which is an improvement of about 11.10% compared with the original algorithm; at the same time, it reached about 84.94% in the Avg.MPA metric, which is an improvement of about 10.84% compared with the original algorithm. These results demonstrate the excellent performance of the improved segmentation network in the leaf segmentation task.

The application of the SAM model in tobacco leaf image segmentation yields promising results for the first time. This achievement not only contributes to the advancement of the tobacco field but also validates the potential of the SAM model within this domain. It serves as a foundation for further research and development efforts. Future work should focus on further enhancing the SAM model. Model compression techniques such as pruning, quantization, and separate convolution can be employed to reduce the model size and computational complexity, thereby improving inference speed. Additionally, the specific algorithms within the SAM model can be refined and optimized. This includes designing more effective attention mechanisms and incorporating contextual information to enhance the accuracy and efficiency of image segmentation. Furthermore, parallel computing and asynchronous inference techniques can be explored to distribute computational tasks among multiple computing units or parallel processors. This approach can significantly accelerate the inference process of image segmentation. By addressing these aspects, future enhancements to the SAM model will enable improved performance and broader applicability in the field of tobacco leaf image segmentation.

REFERENCES

[1] H. Zhang, H. Zhou, J. Zheng, Y. Ge, and Y. Li, "Research progress and prospect in plant phenotyping platform and image analysis technology," *Trans. Chin. Soc. Agricult. Machinery*, vol. 51, no. 3, pp. 1–17, 2020.

[2] N. Nikbakhsh, Y. Baleghi, and H. Agahi, "A novel approach for unsupervised image segmentation fusion of plant leaves based on *g*-mutual information," *Mach. Vis. Appl.*, vol. 32, no. 1, pp. 1–12, Jan. 2021.

[3] F. Fiorani and U. Schurr, "Future scenarios for plant phenotyping," *Annu. Rev. Plant Biol.*, vol. 64, pp. 267–291, Apr. 2013.

[4] C. Li, R. Adhikari, Y. Yao, A. G. Miller, K. Kalbaugh, D. Li, and K. Nemali, "Measuring plant growth characteristics using smartphone based image analysis technique in controlled environment agriculture," *Comput. Electron. Agricult.*, vol. 168, Jan. 2020, Art. no. 105123.

[5] T. Van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Comput. Electron. Agricult.*, vol. 177, Oct. 2020, Art. no. 105709.

[6] M. Minervini, A. Fischbach, H. Scharr, and S. A. Tsaftaris, "Finely-grained annotated datasets for image-based plant phenotyping," *Pattern Recognit. Lett.*, vol. 81, pp. 80–89, Oct. 2016.

[7] D. Ward and P. Moghadam, "Scalable learning for bridging the species gap in image-based plant phenotyping," *Comput. Vis. Image Understand.*, vol. 197, Aug. 2020, Art. no. 103009.

[8] H. Zhang, H. Li, N. Chen, S. Chen, and J. Liu, "Novel fuzzy clustering algorithm with variable multi-pixel fitting spatial information for image segmentation," *Pattern Recognit.*, vol. 121, Jan. 2022, Art. no. 108201.

[9] J.-M. Pape and C. Klukas, "Utilizing machine learning approaches to improve the prediction of leaf counts and individual leaf segmentation of rosette plant images," in *Proc. CVPPP Workshop*, vol. 3, 2015, pp. 1–12.

[10] G. Viaud, O. Loudet, and P. H. Cournède, "Leaf segmentation and tracking in *Arabidopsis thaliana* combined to an organ-scale plant model for genotypic differentiation," *Frontiers Plant Sci.*, vol. 7, p. 2057, Jan. 2017.

[11] X. Yin, X. Liu, J. Chen, and D. M. Kramer, "Multi-leaf alignment from fluorescence plant images," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2014, pp. 437–444.

[12] B. Romera-Paredes and P. Torr, "Recurrent instance segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 312–329.

[13] M. Ren and R. S. Zemel, "End-to-end instance segmentation with recurrent attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 293–301.

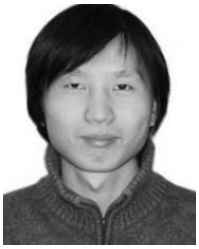
[14] D. Kuznichov, A. Zvirin, Y. Honen, and R. Kimmel, "Data augmentation for leaf segmentation and counting tasks in rosette plants," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–10.
[15] M. A. Khan, T. Akram, Y. D. Zhang, and M. Sharif, "Attributes based skin lesion detection and recognition: A mask RCNN and transfer learning-based deep learning framework," *Pattern Recognit. Lett.*, vol. 143, pp. 58–66, Mar. 2021.
[16] J. Wu and X. Wei, "Omnidirectional gradient and its application in stylized edge extraction of infrared image," in *Proc. Int. Conf. Image Process., Comput. Vis. Mach. Learn. (ICICML)*, Oct. 2022, pp. 98–102.
[17] Y. Zhang, Q. Fan, F. Bao, Y. Liu, and C. Zhang, "Single-image super-resolution based on rational fractal interpolation," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3782–3797, Aug. 2018.
[18] R. Li, S. Zheng, C. Duan, Y. Yang, and X. Wang, "Classification of hyperspectral image based on double-branch dual-attention mechanism network," *Remote Sens.*, vol. 12, no. 3, p. 582, Feb. 2020.
[19] X. Zhang, G. Sun, X. Jia, L. Wu, A. Zhang, J. Ren, H. Fu, and Y. Yao, "Spectral–spatial self-attention networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5512115.
[20] C. Zhang, L. Liu, Y. Cui, G. Huang, W. Lin, Y. Yang, and Y. Hu, "A comprehensive survey on segment anything model for vision and beyond," 2023, *arXiv:2305.08196*.

WEIZHENG ZHANG received the B.S. degree in communication engineering from the Zhongyuan University of Technology, China, in 2005, the M.S. degree in signal and information processing from the Shandong University of Science and Technology, China, in 2012, and the Ph.D. degree in agricultural electrification and automation from Zhejiang University, China, in 2016. He is currently a Lecturer with the College of Computer Science and Technology, Zhengzhou University of Light Industry. His current research interests include machine learning, image processing, and digital agriculture.

YUEFENG WANG received the B.S. degree in software engineering from Anyang Normal University, China, in 2021. He is currently pursuing the master's degree with the College of Computer Science and Technology, Zhengzhou University of Light Industry. His current research interests include deep learning and image processing.

GUANGCAI SHEN received the M.S. degree in tobacco science and engineering from Henan Agricultural University, China, in 2010. He is currently a Senior Agronomist with Yunnan Tobacco Company, Baoshan Branch. His current research interest includes tobacco cultivation.

CANLIN LI received the B.S. degree in computer science from the National University of Defense Technology, China, in 1998, the M.S. degree in computer science and technology from Zhejiang University, China, in 2004, and the Ph.D. degree in computer science and engineering from Shanghai Jiao Tong University, China, in 2010. He is currently an Associate Professor and a M.S. Tutor with the College of Computer Science and Technology, Zhengzhou University of Light Industry. His research interests include image processing, pattern recognition, artificial intelligence, and visual media computing.



MENG LI received the B.S. degree in plant protection from Henan Agricultural University, China, in 2004, the M.S. degree in agricultural entomology and pest control from Northwest A&F University, China, in 2007, and the Ph.D. degree in pest management from Zhejiang University, China, in 2010. He is currently a Lecturer with the College of Tobacco Science and Engineering, Zhengzhou University of Light Industry. His current research interests include tobacco biotechnology, green prevention, and control of diseases and pests.

YINGCHENG GUO received the B.S. degree in agriculture from Yunnan Agricultural University, China, in 2007. He is currently a Senior Agronomist with Baoshan Branch, Yunnan Tobacco Company. His current research interests include tobacco cultivation and pest management.