IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 29 December 2023, accepted 4 February 2024, date of publication 22 February 2024, date of current version 1 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3368911

RESEARCH ARTICLE

An Improved Transformer Network With Multi-Scale Convolution for Weed Identification in Sugarcane Field

CUIMIN SUN[1], MENGHUA ZHANG[1], MUCHEN ZHOU[2], AND XINGZHI ZHOU[1]
[1]School of Computer and Electronic Information, Guangxi University, Nanning 530000, China
[2]School of Mechanical Engineering, Guangxi University, Nanning 530000, China

Corresponding author: Cuimin Sun (cmsun@gxu.edu.cn)

This work was supported by Guangxi University Young and Middle-Aged Teachers Basic Research Ability Improvement Project under Grant 2023KY0020.

ABSTRACT Automated weeding equipment is urgently needed to deal with weeds in farmlands in the context of the rapid development of Intelligent Agriculture. A system for accurately identifying crops and weeds in images is crucial component of automated weeding equipment. However, in the field environment, crops and weeds grow intertwined, and weeds are very similar to sugarcane leaves, making it difficult to accurately segment crops, weeds, and their boundaries from images. In this paper, we proposed a novel network that fully utilizes low-level semantic information to accurately segment crops and weeds in images, improving the accuracy of crop and weed segmentation while reducing the need for training weight parameters and improving speed in the prediction stage. Specifically, we made three important modifications for crop and weed identification. First, a Multi-scale Feature Extraction and Fusion module (MFEF) was designed to capture abundant low-level semantic feature information. Afterward, we introduce a Global Response Normalization (GRN) block to select more useful feature information. Finally, a series of residual attention transformer layers are designed to transmit the long-range dependency information extracted between layers. Numerous experimental results confirmed that our proposed network achieved excellent performance in segmenting sugarcane and weed images. Specifically, (1) the mean accuracy and, Mean Intersection of Union (MIoU) reached 96.97% and 94.13%, respectively, (2) the training parameters of the model have been reduced by more than 25%, improving the Frames Per Second (FPS) value of the prediction process, and (3) is also effective on the publicly available BoniRob Dataset, indicating that the proposed model has considerable generalization ability. This study provides an accurate weed identification map and has reference significance for subsequent systems as mechanical weeding equipment.

INDEX TERMS Weeds identification, semantic segmentation, transformer, Segformer, precision agriculture.

I. INTRODUCTION

With the continuous improvement of agricultural productivity, inefficient production methods are gradually being replaced. Precision Agriculture (PA) is an important component of high-quality agricultural development [1]. Precision Agriculture, also known as site-specific weed management, aims to reduce the cost of herbicides, improve weed control methods, and avoid environmental pollution [2]. Sugarcane is an important economic and energy crop, and also the main raw material for sucrose production [3]. However, the wild growth of weeds in the field environment competes with crops for resources such as water, nutrients, and sunlight, seriously affecting crop growth and ultimately having a negative impact on crop yield [4]. Generally, weeding methods include manual weeding, chemical weeding, and mechanical weeding. Manual weeding is labor-intensive, inefficient and can easily cause heatstroke among farmers

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

in hot weather. Pesticides used in chemical weeding cause grave damage to farmlands and the surrounding ecological environment. Based on the above reasons, it is necessary to find an effective, safe, and low-cost weed control method. Relatively speaking, mechanical weeding is a more suitable weeding method for precision agriculture development in the field. Therefore, there is an urgent need to develop a method that can accurately segment crops and weeds from images [5].

Computer vision was first developed for weed identification. Several methods have been developed by researchers for weed segmentation [6], [7], [8], [9]. The process of using computer vision for weed segmentation includes image acquisition, preprocessing, segmentation, feature extraction, and classification [10]. Various devices are used to capture raw images in the field, including thermal cameras, spectral cameras, and remote sensing devices [11], [12], [13], [14], [15]. When images are collected using these devices, they are susceptible to natural circumstances, such as illumination conditions, soil wetness, or drought. In addition, they are relatively expensive for ordinary people. In terms of the way images are collected, the collected images include airborne remote sensing images and ground-based sensing images [16], [17], [18]. However, in the process of data collection in the sensing images, the data collection points are far away from crops and weeds, and the shooting angle is single, resulting in some weeds being ignored and low spatial resolution of the images. This is not conducive to precise weeding of farmlands. So, we use low-cost digital cameras to capture ground-based sensing images under natural conditions. In the feature extraction process, based on the obtained dataset, some feature combinations that can distinguish crops and weeds are extracted, such as shape, color, texture, and spectral features. Then, the extracted feature matrix is fed into the machine learning algorithm to identify specific classifications. Machine learning algorithms include decision trees, support vector machines, and Bayesian decision theory. Zou et al. [19] proposed an algorithm that combined color and texture features with support vector machines. The algorithm first extracts six colors and five texture features, and then uses a support vector machine as the final classifier. Finally, the segmentation accuracy was 90%. Golzarian et al. [20] combined three types of features, including color, texture, and shape, which were then reduced to three descriptors using Principal Component Analysis (PCA). The results showed accuracies of 88% and 85% in the differentiation of ryegrass and brome grass from wheat, respectively. Under ideal scenarios and during specific phases of plant development, these approaches yield accurate segmentation results. However, the accuracy of these approaches can be affected by multiple factors, such as the type of plants, distribution of weeds, diverse lighting conditions, overlap of crop and weed leaves, and growth stages of the plants [21], as these factors are constantly changing in the actual field. Therefore, it is necessary to develop an economical, efficient, and robust algorithm for this complex environment [22].

In recent years, Convolutional Neural Networks (CNNs) have been rapidly developed and achieved excellent results in various fields [23], [24], [25], [26], [27], [28], [29]. It can automatically learn multi-dimensional feature information with a distinguishing degree between crops and weeds from input images. Owing to its advantages, CNNs are also widely used in the agricultural field to solve various practical problems including weed identification [30], [31], [32], [33]. Zou et al. [34] proposed a U-Net variant network for

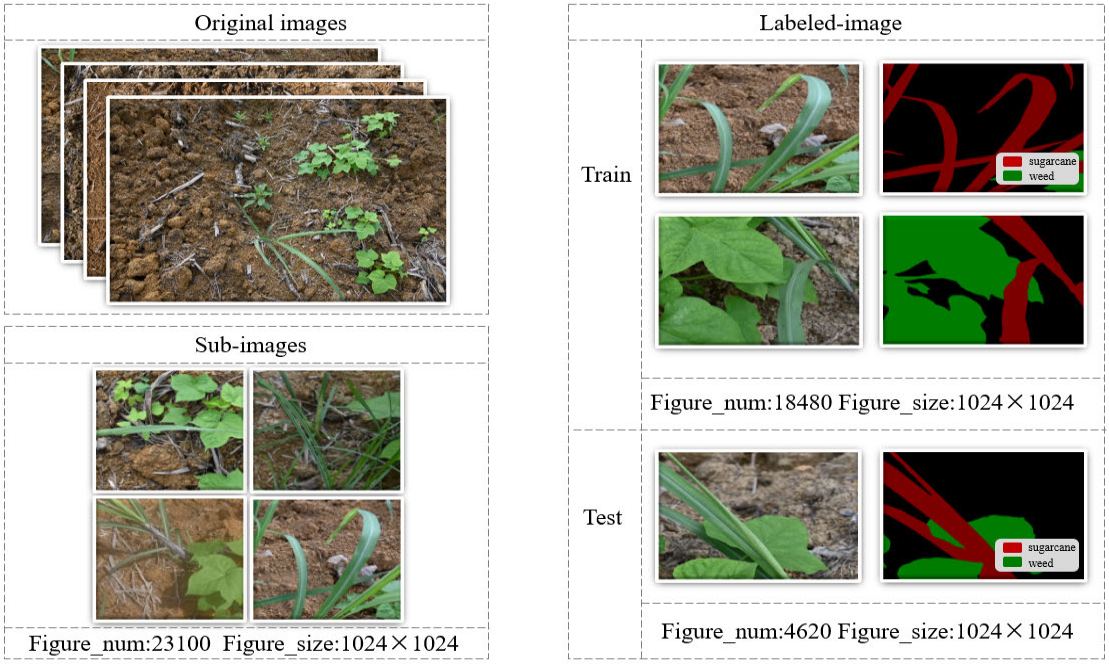


FIGURE 1. Data collection in sugarcane field and weed dataset construction.

segmenting wheat and weeds on digital images. Overlapping dilation convolutions are added to the network to obtain continuous new information on the features of larger receptive fields. Dilated convolutions can cause chessboard. The network yielded a mean intersection over union (MIOU) of 88.98%. Das and Bais. [35] constructed a novel network, named DeepVeg, in which the pyramid pooling module was introduced into the encoder and decoder network to extract multi-level resolution features. The MIOU and accuracy of the DeepVeg model are 0.76 and 0.97, respectively. Jin et al. [36] introduced the CBAM module into the Mask R-CNN network to focus on effective features, suppress invalid features, improve the sensitivity of the model to weed boundaries, and improve the efficiency of model feature learning. The Intersection over Union (IoU) and mean pixel Accuracy (Acc) reached 50% and 94.8% for the beet and miscellaneous vegetable datasets, respectively. CNNs have achieved good results in weed segmentation. However, owing to the intrinsic locality of CNNs, they generally demonstrate limitations in explicitly modeling long-range dependencies.

Recently, models based on transformers have been rapidly developed and are widely used to handle computer vision tasks. In convolutional neural networks, due to the fixed size of the convolutional kernel, there is a local receptive field, and the converter can use attention blocks to capture long-range dependencies in the data, which can solve the problem of crops and weeds being very similar and difficult to distinguish in the local receptive field. Jiang et al. [37] explored the training results of three different transformer models, including Swin-transformer, Segformer, and Segmenter. The results showed that Swin-transformer had the best performance, but had the highest number of training parameters. In contrast, Segformer has performance close to its performance, but the number of training parameters is only one-seventh that of it. Although the transformer can capture long-range dependency information, its ability to handle detailed boundary information is still insufficient. So, in this article, we explore a method of combining convolutional modules with transformer modules. First, the convolutional module is used to obtain more local feature information, and then the transformer is used to capture long-range dependency information between pixels.

In conclusion, the aim of this study is to present an enhanced Segformer model that can achieve precise segmentation of crops and weeds from RGB images. Specifically, this study focuses on three aspects. First, in order to obtain multi-scale local information, we designed a multi-scale feature extract and fusion module with multiple depth-wise convolution paths. Afterward, we introduce the Global Response Normalization (GRN) block to extract more useful feature information. In addition, we embedded the remaining connections into the transformer layer of Segformer to accelerate the transmission of long-range dependency information between the layers.

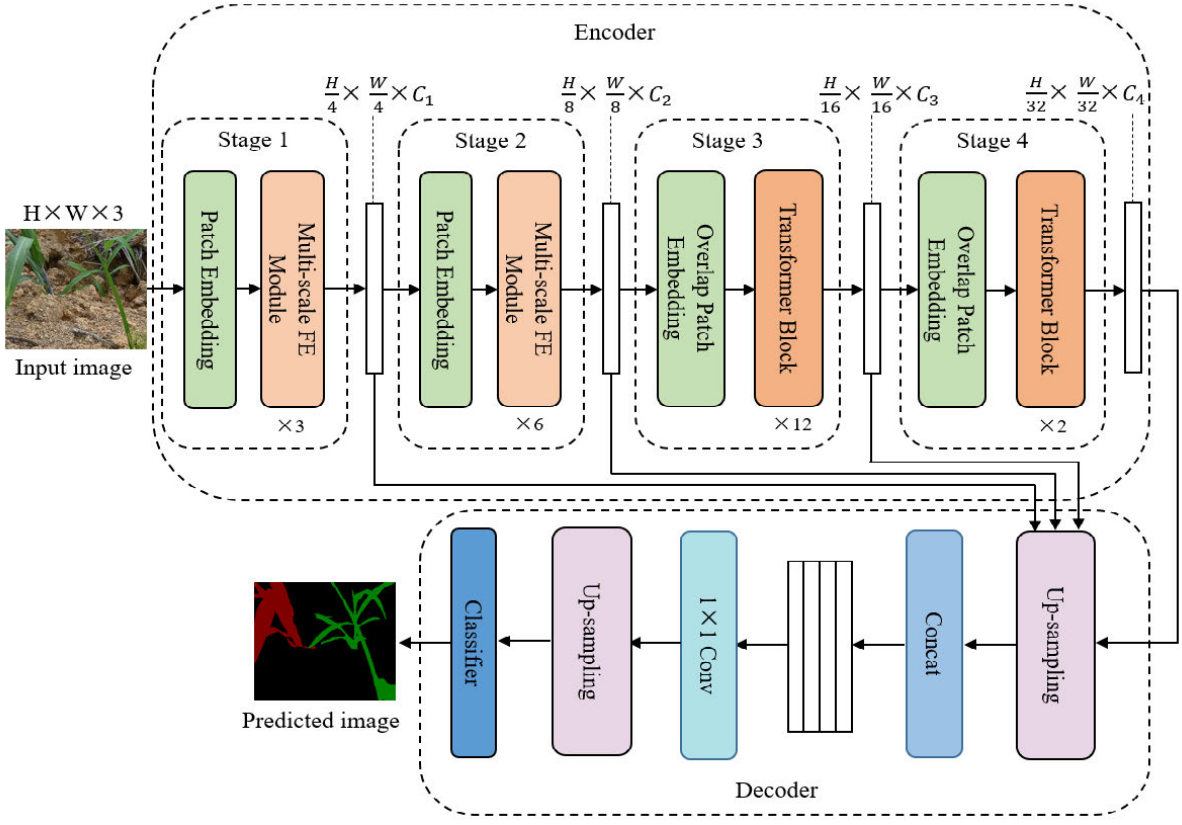


FIGURE 2. Structure of the proposed network.

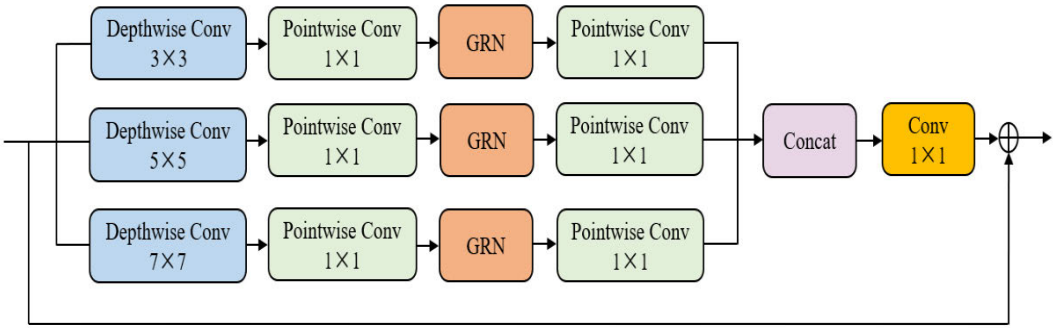


FIGURE 3. Multi-scale feature extraction and fusion module.

II. RELATED WORKS

Recently, models based on transformer have attracted great attention in the field of computer vision. However, there are few research achievements applied to the identification of crops and weeds. In this section, we briefly overview the Transformer model and its applications in computer vision, and review existing weed recognition methods based on convolutional neural networks.

A. THE EVOLUTION OF MODELS BASED ON TRANSFORMER

Transformer was first proposed in the field of Natural Language Processing (NLP). It calculates the correlation between each patch based on the self-attention mechanism. Because the input of the transformer is multiple patches, while the input of the convolutional layer is a single image, it is difficult to apply the transformer to the field of computer vision. Carion et al. [38] use ResNet as the backbone of the network, gradually reducing the input image size, and then feeding it into an encoder-decoder composed of transformers. The experimental results indicate that the transformer can achieve good results in the field of object detection. Subsequently, Ranftl et al. [39] proposed a universal vision backbone based on the transformer, which can apply various computer vison tasks. Firstly, the input image is divided into multiple patches, and then the patches are input into the transformer encoder. By using skip connections, the up-sampling output of the decoder is fused with the output of the corresponding layer in the encoder to generate a fine-grained prediction Liu et al. [40] proposed a Swin-transformer block, which consists of two parts: window multi-head self-attention (W-MSA) and shifted window multi-head self-attention (SW-MSA). W-MSA calculates the self-attention of all pixels within the window. In order to solve the problem of not being able to calculate the self-attention of pixels on the boundaries of the window and other windows, SW-MSA was proposed. The Swin-transformer not only achieves near global attention capability, but also reduces the computational complexity from a square relationship of image size to a linear relationship, greatly reducing computational complexity and improving model inference speed. Xie et al. [41] designed an effective self-attention module, which effectively reduced the computational complexity of self-attention. Specifically, a hyperparameter R is introduced in the effective self-attention module, which down-sampling the input patch by R times, thereby reducing the computational complexity of the self-attention module. Niu et al. [13] proposed a novel HSI-TransUnet, which combines convolutional modules with transformers and can fully utilize the rich spatial nuclear spectral information of drone HSI data, achieving good results.

B. APPLICATION OF CONVOLUTIONAL NEURAL NETWORKS IN THE FIELD OF AGRICULTURE

Due to the unique advantages of convolutional neural networks, they have developed rapidly in the field of agriculture. Numerous research papers have proven its effectiveness in identifying crops and weeds. Next, we provide a brief overview of previous research based on this method.

Training deep neural networks requires a large number of labeled samples, and the labeled samples of semantic segmentation networks require pixel by pixel annotation of images, which consumes a lot of manpower and material resources. Zou et al. [18] modified the U-Net network by fusing more low-dimensional semantic information from the decoder with the feature information of decoder using a skip connection structure, enhancing the segmentation accuracy of the model. In addition, a data augmentation method was proposed to increase the number of pre-trained datasets. The collected images are segmented into weeds using the minimum error segmentation algorithm, and then the weed image is used as the ''foreground'' and other images are used as the ''background'' to synthesize pre-trained images. Kim et al. [42] proposed a multi-task semantic segmentation convolutional neural network. Specifically, the task was divided into two steps to achieve. The first step was to segment images that removed the background but included crops and weeds from the image, and the second step was to segment crops and weeds from the output of the previous step. In addition, different loss functions were used in the two segmentation tasks. Nasiri et al. [43] improved U-Net by adding residual connections to the convolutional layer of the encoder, preserving more detailed information. In order to optimize the problem of data imbalance and small area

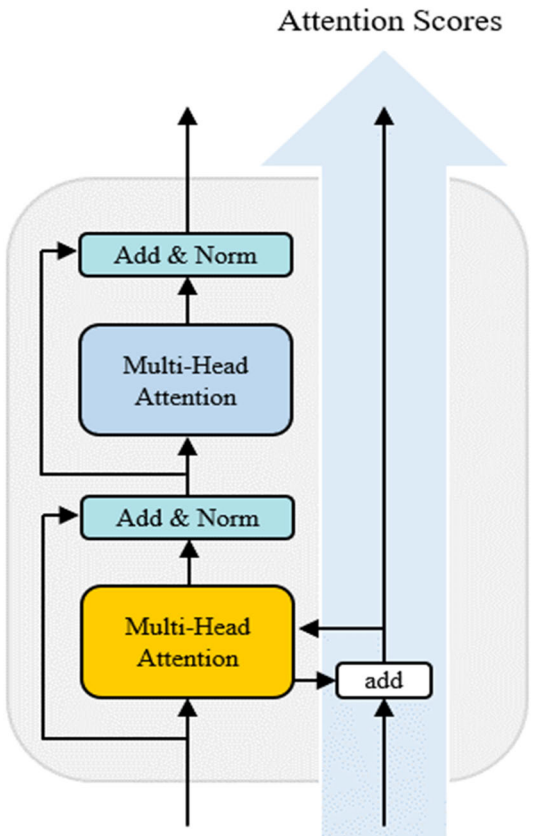


FIGURE 4. Multi-scale feature extraction and fusion module.

precise segmentation, the combination of dice and focal losses is used as the loss function of the model. These methods are all aimed at obtaining more detailed information to optimize precise boundary segmentation.

In this article, we explore the combination of convolutional module and transformer module, utilizing convolutional module to extract rich low dimensional semantic information, and then capturing long-range dependency information with transformer to establish complete boundary information of the target.

III. EXPERIMENTAL DATA AND METHODS

A. DATA ACQUISITION AND PROCESS

In this research, a digital camera (Nikon Z5) was used to collect images of the crops and weeds. In order to obtain sufficient real-scene images of crops and weeds, we collected images in multiple time periods, different weather conditions, and multiple growth stages of sugarcane. Specifically, the images were taken from 9:00 to 18:00 during the period from May 26, 2022, to June 1, 2022; and from 9:00 to 18:00 from July 26, 2022, to July 28, 2022, from 9:00 to 18:00 during the period from September 19, 2022, to September 21, 2022, at Guangxi University Agricultural and Animal Husbandry Industry Development Research Institute in Guangxi Autonomous Region (107°47′, 22°31′). The original resolution of the image we collected was 6016 × 4016 pixels, which was too large and required a GPU device with a large graphics memory for training. Considering the future practical application of equipment parameters and reduction of experimental costs, it is necessary to split the original image into smaller sizes. In order to preserve the boundary information of the images, we use an overlapping cropping method to crop the image into 1024 × 1024 pixels, with overlapping parts in steps of 512 pixels. In total, 23100 images were obtained. All images were manually labeled pixel by pixel using EISeg annotation software [44], and each image took more than an hour. In the annotated image, the background, sugarcane, and weeds were labeled 0, 1, and 2, respectively. Furthermore, their pseudo-colors were black, red, and green, respectively. The training dataset consisted of 80% of the images and their corresponding pixel-level labels, while the remaining 20% was used to test the model. The description of the dataset is shown in Figure 1.

B. THE PROPOSED NETWORK STRUCTURE

In this section, we describe the details of the proposed network, as shown in Figure 2. The transformer module is used to calculate the attention of each pixel and other pixels in an image. This calculation method takes a lot of time in a single image, resulting in a slow training process. In Segformer, an efficient multi-head self-attention module is proposed to reduce the computational complexity of the original transformer module. Specifically, the original calculation of attention required the input of Q, K, and V as the three parameters. The improved method is to down-sample K with hyperparameter R, with R set to 8, 4, 2, and 1 in each of the four stages. That is to say, only calculate the self-attention before the pixels in the patch and the pixel points generated after down-sampling the patch. Good results can be achieved in super-resolution image segmentation. However, some details were lost in the complex environment of sugarcane fields. To address this issue, we have proposed some improvement methods. Concretely, the encoder part consists of a multi-scale feature extraction and fusion module (MFEF), Global Response Normalization (GRN), and residual transformer block (RTB). The MFEF uses multi-scale convolution to capture the feature information of multi-scale receptive fields, playing a role in dynamically changing the size of the convolution kernel, that is to say, dynamically selecting the optimal convolution kernel size that best represents nearby pixels. GRN aims to strengthen the contrast and selectivity of channels. The RTB is designed to transmit the long-range dependency information extracted between layers. As for the decoder part, a semantic segmentation classifier was added on top of the original lightweight decoder.

1) ENCODER STRUCTURE

The encoder part plays an important role in semantic segmentation networks. Its main responsibility is to extract rich abstract feature information from the input image layer by layer. Convolutional neural networks typically use fixed-sized convolutional kernels. In order to obtain a large

receptive field, researchers typically use pyramid pooling and dilated convolution. However, the former may overlook some detailed information, whereas the latter may lead to a chessboard effect. Inspired by [45] and [46], we built a multi-scale feature extraction and fusion module, as shown in Figure 3, which adopts a mixed-size convolutional kernel to capture abundant low-level details. First, a 1 × 1 convolution was used to reduce the number of channels in the feature map, thereby reducing the computational complexity. Then, it is split into three branches; the 3 × 3, 5 × 5, and 7 × 7 depth-wise convolutions (dwConvs) are used to increase the multi-scale local information extraction of our network. Finally, we used a 1 × 1 convolution to restore the number of feature map channels to the original number of input channels. Besides, we used deep separable convolutions instead of comm convolutions to reduce computational complexity. In [47], introducing a GRN enhances inter-channel feature competition. GRN improves representation quality by enhancing feature diversity. The results obtained were better than those obtained using SE [48] and CABM [49] modules. Inspired by this, we added this module to our network to enhance feature competition between channels. Actually, there are four stages that are stacked by multiple transformer modules in the original network. A deeper network structure can cause some problems, including slowing down the speed of the network and the vanishing gradient problem during loss back-propagation. To alleviate this issue, we introduced a Residual Attention Layer Transformer in the structure of the stacked transformer layers of the last two stages in the Segformer [41], as shown in Figure 4.

2) DECODER STRUCTURE

In order to fuse the low-level semantic information from encoder in the decoder part of the network structure to make the segmentation boundary more accurate, we made a modification on the basis of traditional network encoders, integrating the outputs of the four stages of the encoder into the decoder. First, the hierarchical feature information generated by the encoder is up-sampled to the size of $\frac{W}{4} \times \frac{H}{4}$ using bilinear interpolation. Then, we concatenated feature map in channel direction. Next, the fused feature maps are feed into a 1 × 1 convolution layer to unify the channel dimension, and a 4 times up-sampling operation to the size of H × W. Finally, the up-sampled fused feature map is processed through another MLP layer to predict the classes result with a resolution of H × W×N_{cls}, where N_{cls} is the number of classes.

C. PARAMETER CONFIGURATION DURING NETWORK TRAINING

The hardware environment was an Intel Xeon W-2235 CPU, 64 GB memory, and NVIDIA GeForce RTX 3090 with 24GB of video memory. The software environment was Windows 10, CUDA 11.3, Python 3.8, and PyTorch 1.11.0, as shown in Table 1.

TABLE 1. Detailed information of the experimental platform.

Name	configuration
CPU	Inter Xeon W-2235
GPU	NVIDIA RTX3090
CUDA	11.3
cuDNN	V11.3.58
Python	3.8

TABLE 2. Hyperparameter initialization settings.

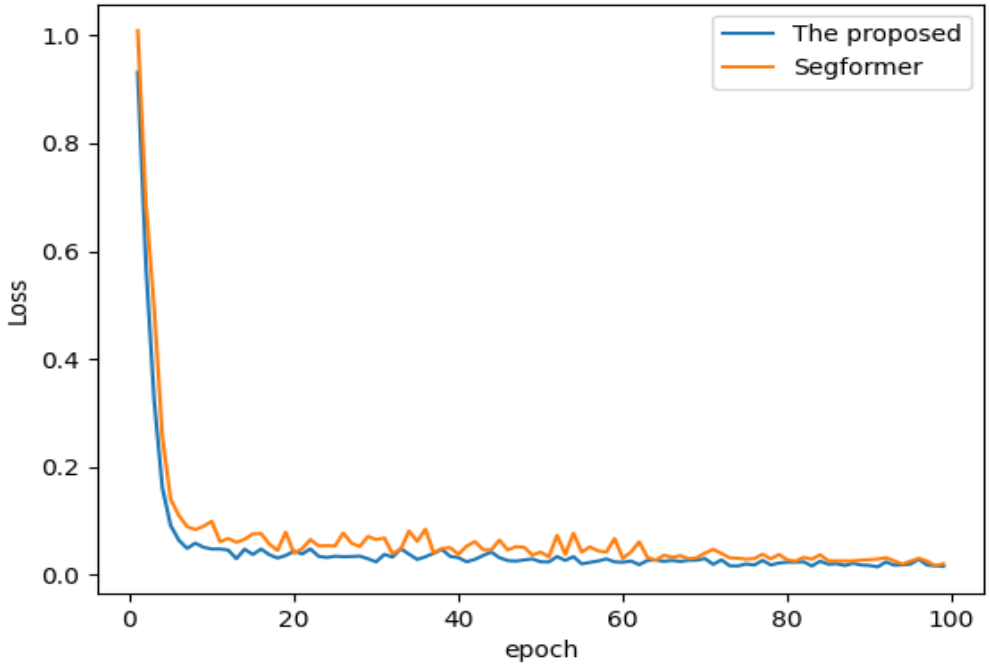
Hyperparameter	value
optimizer	AdamW
learning rate	10^{-5}
weight decay	10^{-4}
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$
batch size	8
training epochs	100
warmup epochs	5

In this section, many parameters were initialized and set, as shown in Table 2. AdamW is used as the optimizer of the training network, the range of the β parameter is 0.9-0.999, the weight decay was 10^{-4}. The initial learning rate was 10^{-5}, the attenuation function of the learning rate is cosine function, the batch size is 8, and the training epoch is 100. In the initial stage of training, it was difficult for the model to maintain parameter stability. Warm-up technology has been widely used to accelerate the convergence rate of the network and reduce instability training [50]. The Cross-entropy loss function is used as the network loss function.

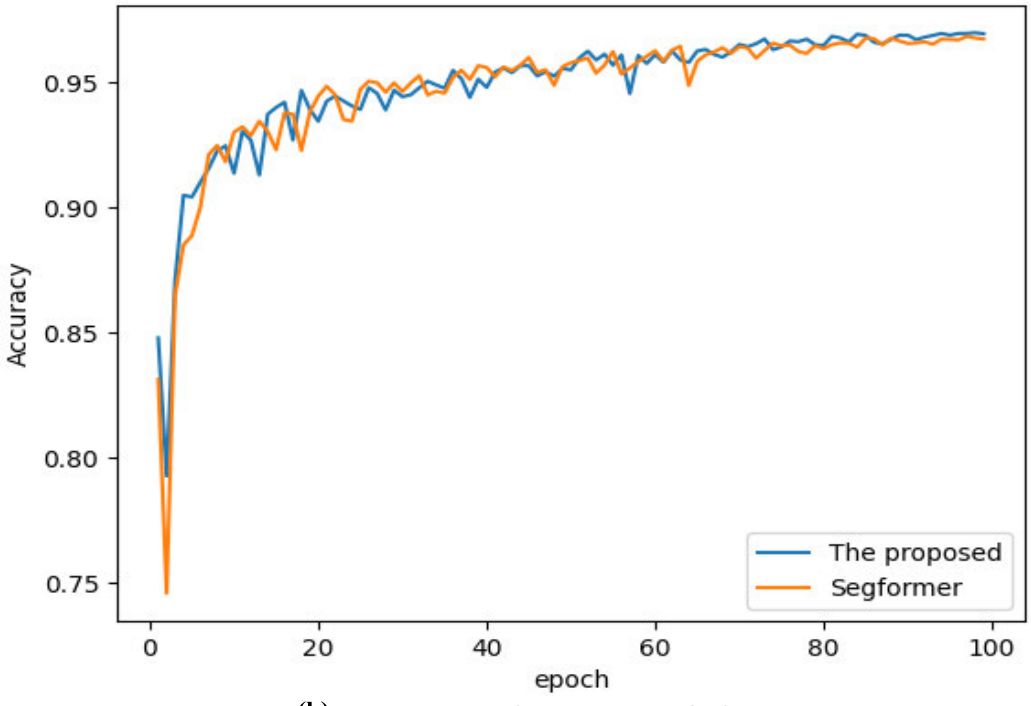
To prevent overfitting of the model, data augmentation was performed on the input image. There are many methods of data augmentation, such as rotation, mirror image, and increasing noise [51]. We used scaling jitter, horizontal flipping, and color jitter to enhance the original image, thereby increasing the diversity of the input image and improving the robustness of the network. Specifically, scaling jitter is the process of randomly resizing to within the range of [0.5, 2.0], followed by randomly cropping out the 1024 × 1024 sized image. Secondly, flip the input dataset horizontally with a probability of $p = 0.5$. Finally, for color jitter, first convert the image to the HSV color model, then change the saturation and hue of the image with a probability of $p = 0.5$, and then convert the image to the BGR color model. In addition, dropout is also introduced in the encoder to prevent overfitting of the training result.

D. EVALUATION METRICS

We used five metrics to evaluate the performance of the proposed network, including pixel Accuracy (Acc), Precision (Pr), Recall (Re), Intersection over Union (IoU), and Fscore,



(a) Loss graphs of the proposed and Segformer



(b) Accuracy graphs of the proposed and Segformer

FIGURE 5. Accuracy and loss curves of the proposed and Segformer.

as shown in formulas (1)-(5).

$$\text{Acc} = \frac{\sum TP + \sum TN}{\sum TP + \sum TN + \sum FP + \sum FN} \times 100\% \quad (1)$$

$$\text{Pr} = \frac{\sum TP}{\sum TP + \sum FP} \times 100\% \quad (2)$$

$$\text{Re} = \frac{\sum TP}{\sum TP + \sum FN} \times 100\% \quad (3)$$

$$\text{IoU} = \frac{\sum TP}{\sum TP + \sum FN + \sum FP} \times 100\% \quad (4)$$

$$F_{score} = \frac{2 \times \text{Pr} \times Re}{\text{Pr} + Re} \quad (5)$$

where TP, TN, FP, and FN are respectively true positive, true negative, false positive, and false negative. In this study, each pixel was divided into one of three classes (soil, sugarcane

and weed). So, the Mean Intersection over Union (MIoU) is the average of Intersection over Union for three classes.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. ACCURACY EVALUATION RESULTS

Figure 5 shows the training loss and accuracy graphs for the proposed model and Segformer. As shown in Figure 5(a), the loss graphs show a rapid decrease in the loss values during the first 10 epochs. Our model gradually stabilized after 58 epochs, surpassing 65 epochs of Segformer, and eventually converged to a very small value. This indicates that our model is effective and has strong feature extraction capability stability. In addition, throughout the entire training process, the loss function value of the proposed model was less than that of the Segformer. In Figure 5(b), the accuracy graphs converge to a sufficiently large value as the number of training epochs increases. Overall, the accuracy function curves of the two networks are very similar, with a final accuracy exceeding 96%. To be precise, the accuracy value of the proposed network changed slightly less than that of the Segformer, and the final accuracy was slightly higher. From table 3, it can be seen that our network achieved MIOU, Acc, Fscore, Pr, and Re values of 94.13%,96.97%, 96.88%, 97.01%, and 96.97, respectively. This indicates that our network has good segmentation results in crop weed segmentation and exceeds those of the benchmark network.

In addition, we used ROC-AUC curves to specifically analyze the accuracy of each category predictions. ROC-AUC curve is a tool used to evaluate the accuracy of classification models, where the horizontal axis represents false positive rate (FPR), the vertical axis represents true positive rate (TPR), and the area under the curve is AUC. The closer its value is to 1, the better the classification accuracy. To enable it to be used for the analysis of the results of multi-classification semantic segmentation, when calculating the FPR and TPR of each category, we set the current category in the label to 1 and other categories to 0. The ROC-AUC curve is shown in Figure 7. From the graph, it can be seen that the AUC values of background, sugarcane, and weed are 0.9989, 0.9974, and 0.9983, respectively. Overall, all three values are close to 1, indicating that the accuracy of the current model segmentation results is sufficient.

Finally, in order to more intuitively display the segmentation performance of our network in each category and the distribution of error segmentation results, we used the confusion matrix to visually analyze the segmentation results, as shown in Figure 6. In the Confusion matrix, each row represents the true category, each column represents the prediction category, and the value on the diagonal represents the accuracy of the prediction category. It is obvious that in the prediction results of each category, there is error data for predictions of other categories. However, in the prediction results of each category, there are errors in predicting data for other categories. In the background of the picture, there are some straws, dried leaves, etc., which can lead to incorrect predictions. In the picture, there are some weed texture features that are very similar to crops, which can lead to incorrect predictions between crops and weeds. In addition, at the tip of crop and weed leaves, due to their very slender and weak features, there will be a small amount of predicted background data, as shown in Figure 6.

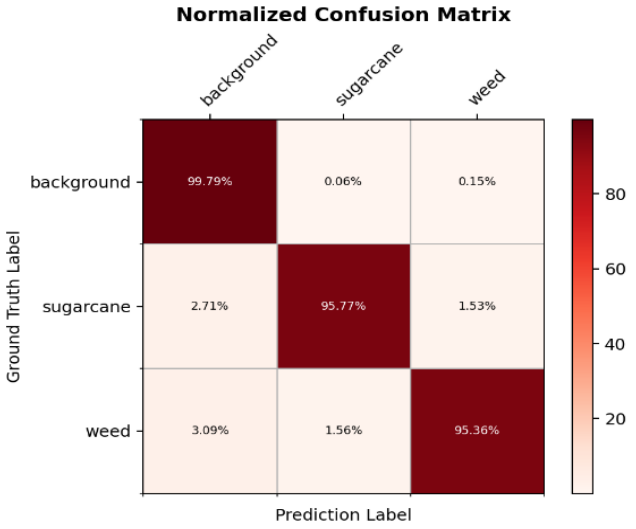


FIGURE 6. Network segmentation results in confusion matrix.

B. COMPARISON BETWEEN THE PROPOSED AND OTHER NETWORKS

Some comparative experiments were conducted, and we compared the performance between the proposed model and seen from Table 5 that the proposed model exhibits the highest performance when compared with the state-of-the-art model. On the test set, the MIOU, Acc, Fscore, precision, and recall of our proposed network were 94.13%, 96.97%, 96.88%, 97.01%, and 96.97%, respectively. U-Net was constructed using at the convolutional neural network without transformer module. Its segmentation accuracy is the worst compared to the other four models. The MIOU, accuracy, Fscore, precision, and recall on the test set were 92.11%, 95.11%, 95.8%, 96.54%, and 95.11%, respectively. Swin-transformer had built a pure transformer backbone network that can be used for various downstream tasks. Its MIOU, accuracy, Fscore, precision, and recall on the test set were 93.41%, 96.01%, 96.54%, 97.09%, and 96.01%, respectively. ConvNeXt V2 is an upgraded version of ConvNeXt [43] that proposes a fully convolutional masked autoencoder framework and a new Global Response Normalization (GRN) based on it. Its MIOU, accuracy, Fscore, precision, and recall on the test set were 93.33%, 96.43%, 96.49%, 96.57%, and 96.43%, respectively.

In addition, we also analyzed each model from three indicators: parameter quantity, computational complexity, and FPS. The size of parameter quantity of model reflects the complexity of model and the memory space required for training the model. FPS represents the number of images that

TABLE 3. Performance comparison table of the proposed network and Segformer.

Method	MIOU (%)	Acc (%)	Fscore (%)	Pr (%)	Re (%)
The proposed	94.13	96.97	96.88	97.01	96.97
Segformer	93.77	96.76	96.74	96.52	96.76

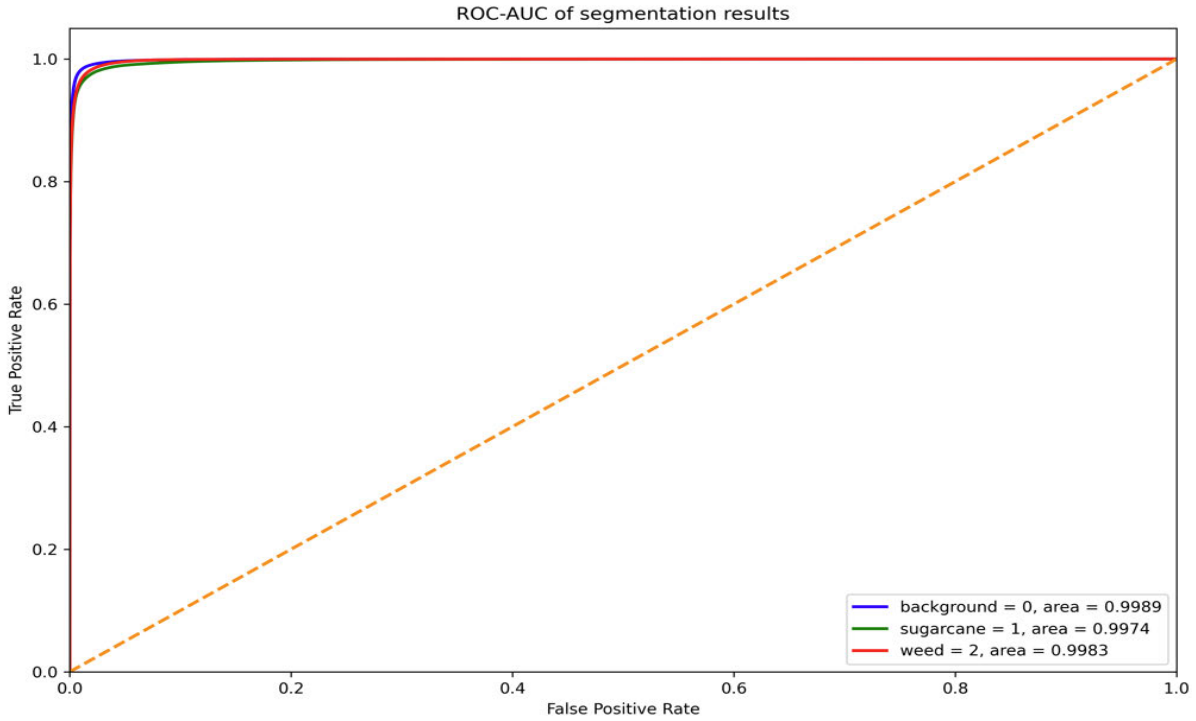


FIGURE 7. ROA-AUC Crop and weed identification results.

TABLE 4. Comparison of relevant indicators of different models.

Method	Params (MB)	GFLOPs	FPS
baseline	47.3	117.9	13.31
U-Net	29.06	810.3	10.42
Swin-transformer	59.83	798.52	10.90
ConvNeXt V2	83.76	853.77	9.18
The proposed	35.46	208.59	13.81

model can process per second during the prediction phase. From Table 4, it can be seen that the network composed of convolutional modules has a relatively large computational load, reaching over 800GFLOPS, and the prediction speed is also relatively slow. Our network is relatively small in terms of parameters, reducing it by 25% compared to baseline. Greatly saving memory requirements in practical applications. In addition, they also achieved the best results in predicting speed. It has to be said that although the computation of our network has increased, it only affects the time required for the training process. We deem that it is acceptable under improving the performance of other indicators.

In Figure 8, the segmentation results of five network images in complex environments are shown, including the situation of leaf crossing and occlusion of sugarcane and gramineous weeds, and sugarcane and broadleaf weeds. From the results, it can be seen that our proposed network improves the ability to identify weeds. More accurate boundary recognition for sugarcane and weeds.

Although good results were achieved in the experiment, there are also some issues that need to be optimized. In actual fields, there are weeds that are very similar to sugarcane, which can affect the accuracy of the model. It is necessary to train the model model with images of large weeds surrounding the sugarcane scene to improve the accuracy of identifying similar weeds. In addition, the model has a relatively large number of parameters and the input images are also large, which makes it impossible to set a larger batch size to train the model and a longer training time.

C. TESTING ON BONIROB DATASET

To verify the universality of our network, we performed additional experiments on the BoniRob Dataset [52]. The whole BoniRob Dataset records the growth records of a sugar beet farm near Bonn in Germany for the past three months, including all data recorded three times a week. We used a dataset

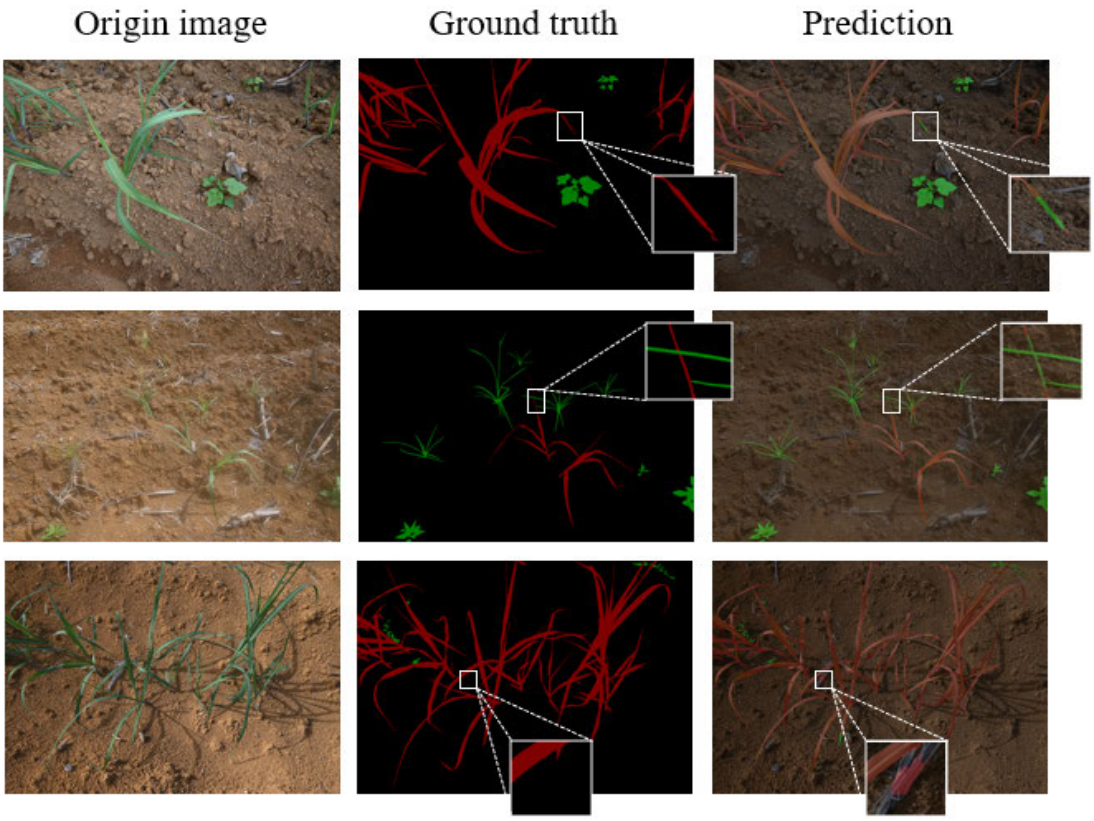


FIGURE 8. Crop and weed identification results with the proposed method.

TABLE 5. Performance comparison for different models.

Method	MIOU (%)	Acc (%)	Fscore (%)	Pr (%)	Re (%)
baseline (Segformer)	93.77	96.76	96.74	96.52	96.76
U-Net	92.11	95.11	95.8	96.54	95.11
Swin-transformer	93.41	96.01	96.54	97.09	96.01
ConvNeXt V2	93.33	96.43	96.49	96.57	96.43
The proposed	94.13	96.97	96.88	97.01	96.97

consisting of 1865 images with a size of 1296 × 966 pixels. The training dataset consisted of 80% of the images and their corresponding pixel-level labels, whereas the remaining 20% were used to test the model. The experimental results indicate that our model can perform well on publicly available datasets and outperform other models, as shown in Table 5.

Due to our added module structure being able to obtain multi-scale feature information, it has good robustness for different datasets. In the comparative experiment, although BoniRob only had a small dataset, the MIOU also reached 88.7%, which is better than the adaptability of other networks.

D. ABLATION STUDY

In this section, as the proposed network is an improvement on Segformer, ablation experiments are performed to verify the effectiveness of various parts of the network. Specifically, the four networks were trained separately. of each modification. Specifically, this section considers the following models.

1) The original Segformer, which utilizes stacked transformer block as encoder.
   Model-a, which replaces the first two transformer stages of Segformer by the proposed multi-scale feature and fusion module.
2) Model-b, which adds GRN module to the multi-scale feature extraction module based on Model-a.
3) The proposed network, which introduces residual connections in the last two transformer stages based on Model-b.
4) Model-b, which adds GRN module to the multi-scale feature extraction module based on Model-a.
5) The proposed network, which introduces residual connections in the last two transformer stages based on Model-b.

All the above models had the same parameter configuration during training. Table 6 shows that compared to the baseline, our proposed network has slightly improved segmentation

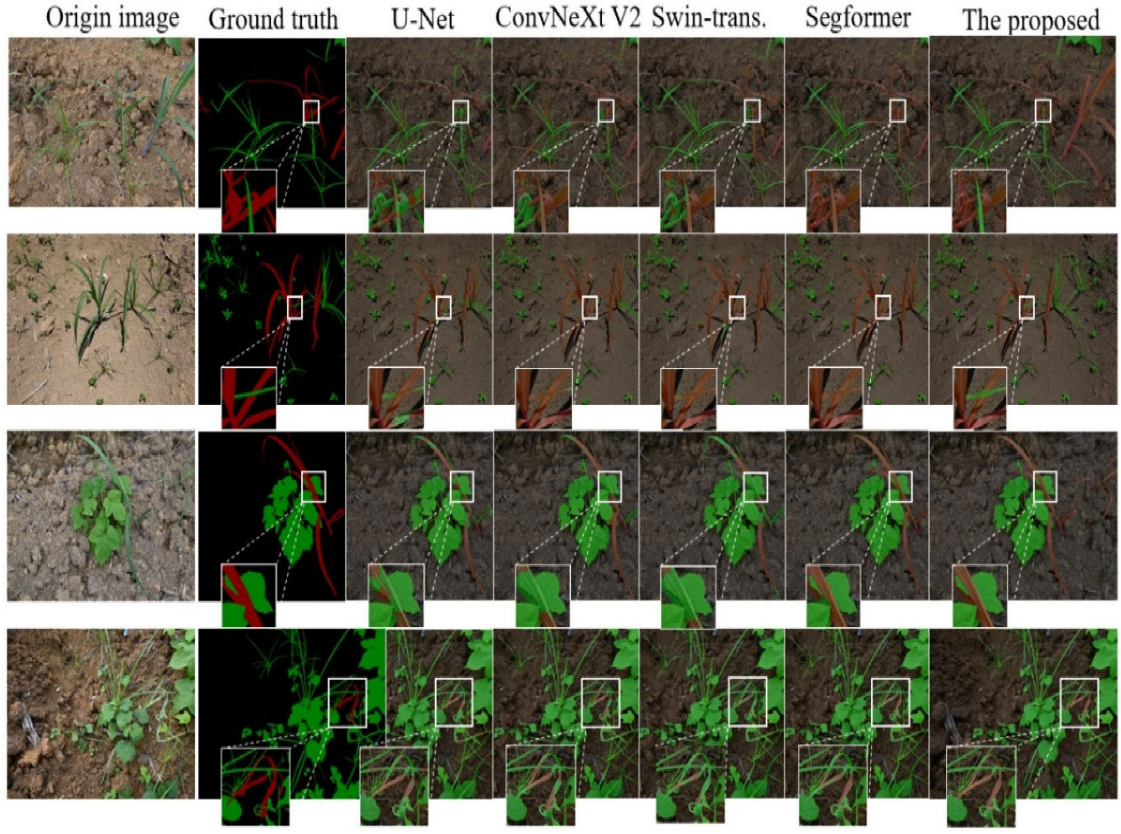


FIGURE 9. Comparison of weed identification performance among various networks.

TABLE 6. Comparison of proposed and different models on BoniRob Dataset.

Method	MIOU (%)	Acc (%)	Fscore (%)	Pr (%)	Re (%)
baseline (Segformer)	84.11	89.56	90.51	91.6	89.56
U-Net	82.11	88.56	90.6	90.11	88.26
Swin-transformer	84.44	90.64	90.77	90.91	90.64
ConvNeXt V2	81.24	87.44	88.55	89.84	87.44
The proposed	88.7	94.07	93.84	93.62	94.07

TABLE 7. Performance comparison for different module configurations.

Method	Params (MB)	MIOU (%)	Acc (%)	Fscore (%)	Pr (%)	Re (%)
baseline	44.6	93.77	96.76	96.74	96.52	96.76
Model-a	35.45	94.07	96.88	96.9	96.93	96.88
Model-b	35.46	94.11	96.92	96.93	96.95	96.92
The proposed	35.46	94.13	96.97	96.88	97.01	96.97

performance. Specifically, the baseline shows the worst predictions with a MIOU of 93.77%. The main reason for this result is that the four transformer stages of the encoder of the Segformer adopt reduction ratios of 8, 4, 2, and 1 time, respectively, to reduce the computational burden of multi-head self-attention. A large reduction ratio can disrupt the texture information of crop and weed and ignore some details that affect the encoder's ability to extract more feature information. As for model-a, it exploits the multi-scale feature extraction and fusion module instead of the first two transformer stages of Segformer which could further refine the texture computation, resulting in a MIOU and Acc increase of 0.3% and 0.12%, respectively. In order to enhance the ability of the feature extraction module to obtain more useful features, we introduced the GRN module to improve MIOU and Acc from 94.07% to 94.11%, and from 96.88% to 96.92%,

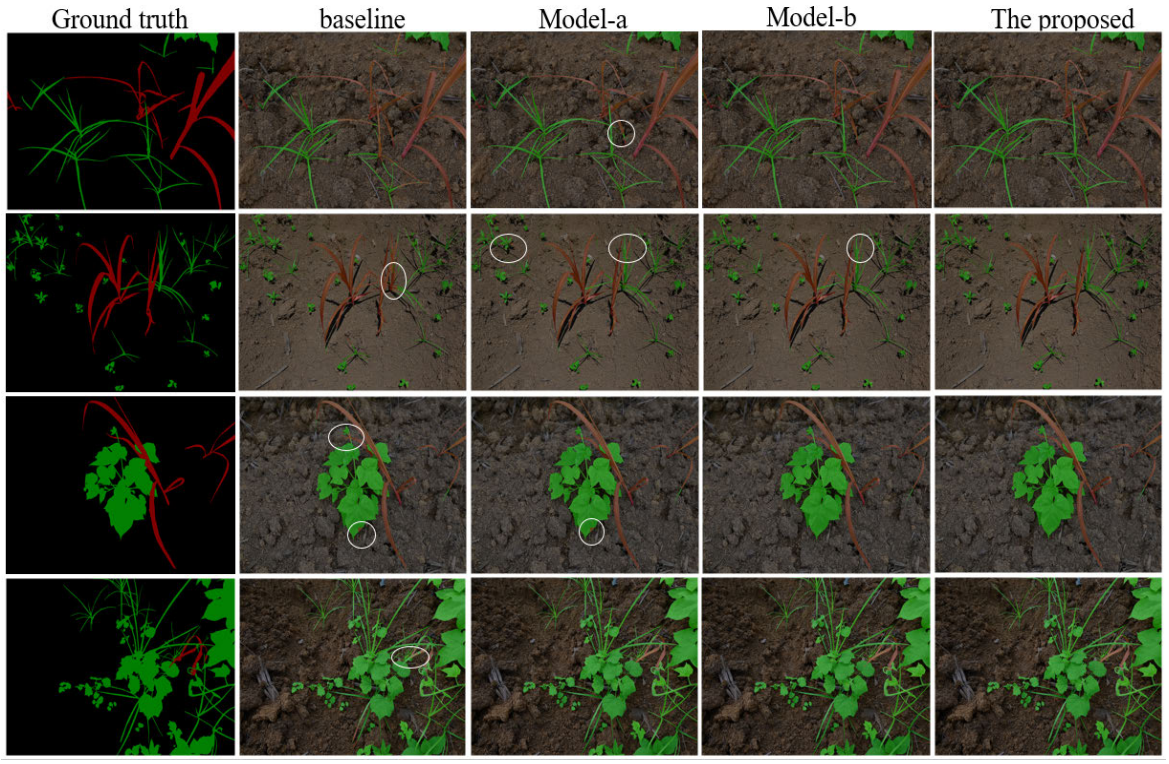


FIGURE 10. Comparison of weed identification performance among various networks.

respectively. Finally, adding residual connections in the transformer layers can accelerate the transfer of long-range dependency information between layers, and increase the MIOU and Acc of model-b from 94.11% and 96.92% to 94.13% and 96.97%, respectively. On the whole, the ablation experiments have proven that each modification is effective, with a total increase of approximately 0.36% in the MIOU and a reduction of 9.14M in parameter quantity compared to the original Segformer. Importantly, the baseline network had poor segmentation performance at the boundaries, intersections, and overlaps between sugarcane and weed. On the contrary, our proposed network achieved good segmentation results. The specific segmentation results are shown in Figure 9.

V. DISCUSSION AND CONCLUSION

The weed identification model is crucial for automatic weed control systems. We choose to use a network based on self-attention mechanism, which has higher accuracy than convolutional neural networks. In the real environment, weeds and sugarcane are not fixed in size, and plant size fluctuates greatly. We need flexible receptive fields to improve the accuracy of weed identification. In this study, we proposed an improved semantic segmentation model for crop and weed segmentation. First, we designed a multi-scale feature extraction and fusion module to extract and aggregate abundant low-level features. Second, we introduce a Global Response Normalization block to enhance inter-channel feature competition so that more useful feature information can be extracted. Thirdly, we embedded residual connections into the stacked Efficient Transformer layers to accelerate the transfer of long-range dependency information between layers alleviate the vanishing gradient issue. In the experimental section, we conducted many experiments by adjusting hyperparameters. When the batch size is increased, the accuracy improvement is minimal, and the cost of memory calls is high, requiring training on more expensive devices. In addition, in cases where the data volume is small, fixing the training epoch and increasing the batch size will increase the risk of overfitting. However, when the batch size is reduced, the accuracy will decrease. So, the current setting is the most suitable. The experimental results demonstrate that our proposed model is effective for crop and weed segmentation with an average accuracy of 96.97%. Compared with Segformer, the proposed model has improved segmentation accuracy by 0.36%, the training speed has been accelerated, and the number of parameters has been reduced by 9.14M. And the proposed network has better segmentation results for the boundary, intersection, and occlusion parts of sugarcanes and weeds. Moreover, our model also achieved good segmentation results on the BoniRob Datasets with an average accuracy of 94.07%. Future research could further reduce the complexity of the model, improve the speed of segmentation, and enable it to run on an embedded system.

VI. LIMITATIONS AND FUTURE WORK

Although our proposed semantic segmentation model achieved good performance in weed recognition tasks, there

are still some shortcomings that need to be further addressed in future work.

First, we only identified crops, weeds, and backgrounds, without specifically identifying the types of weeds. Although this is sufficient for mechanical weed control systems, identifying specific types of weeds is of great help for precise agricultural spraying. It also provides various options for farmers to select agricultural combinations. This is a problem that needs to be addressed in future works. When annotating data, it is necessary to accurately label each weed species, which requires sufficient in-depth research on weeds. Incorrect annotation of images can affect the accuracy of the model in identifying weeds. Therefore, experienced agronomists were required to assist in annotating the data.

Second, transformer network and its variants are extremely sensitive to data and require a large amount of training data. In the agricultural field, relatively few publicly available datasets and they contain a single variety of crops and weeds. To train a universal weed recognition network, a combination of machine learning and deep learning methods are considered for identifying weeds in the future.

Third, our proposed model showed good results for crop and weed segmentation performance. However, the average processing speed during the prediction process is only 13.8 images per second. On the one hand, this is because of the large size and high resolution of our input images; on the other hand, this is because our network has a large number of parameters and computational complexity. Therefore, it is necessary to study the impact of input image resolution on performance, and further optimize the network structure without reducing performance, reduce computational complexity, and increase the FPS in the prediction stage to meet the requirements of weed identification in actual environments.

Fourthly, due to limited data, we only validated the generalization of our network on the BoniRob dataset, and experimental results showed that our network has good generalization performance. In the future, further validation will be conducted on other crop datasets, and our goal is to develop a universal weed identification network.

The above is a solution proposed for the limitation, and the most important issue is that in actual farmland, weeds and crops usually grow together, and they are very similar. So, I will consider using covered object detection networks to detect weeds.

REFERENCES

[1] I. Cisternas, I. Velásquez, A. Caro, and A. Rodríguez, "Systematic literature review of implementations of precision agriculture," *Comput. Electron. Agricult.*, vol. 176, Sep. 2020, Art. no. 105626.

[2] P. Herrera, J. Dorado, and Á. Ribeiro, "A novel approach for weed type classification based on shape descriptors and a fuzzy decision-making method," *Sensors*, vol. 14, no. 8, pp. 15304–15324, Aug. 2014.

[3] Y. G. Ou, M. Wegener, D. T. Yang, Q. T. Liu, D. K. Zheng, M. M. Wang, and H. C. Liu, "Mechanization technology: The key to sugarcane production in China," *Int. J. Agricult. Biol. Eng.*, vol. 6, no. 1, pp. 1–27, 2017.

[4] J. Arroyo, M. Guijarro, and G. Pajares, "An instance-based learning approach for thresholding in crop images under different outdoor conditions," *Comput. Electron. Agricult.*, vol. 127, pp. 669–679, Sep. 2016.

[5] U. Zahra, M. A. Khan, M. Alhaisoni, A. Alasiry, M. Marzougui, and A. Masood, "An integrated framework of two-stream deep learning models optimal information fusion for fruits disease recognition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 3038–3052, 2024.

[6] D.-M. Li, Y.-Z. Wang, and B. Du, "Research on segmentation methods of weed and soil background under HSI color model," in *Proc. 2nd Int. Workshop Knowl. Discovery Data Mining*, Jan. 2009, p. 628.

[7] S. Lavania and P. S. Matey, "Novel method for weed classification in maize field using Otsu and PCA implementation," in *Proc. IEEE Int. Conf. Comput. Intell. Commun. Technol.*, Feb. 2015, pp. 534–537.

[8] F. Lin, D. Zhang, Y. Huang, X. Wang, and X. Chen, "Detection of corn and weed species by the combination of spectral, shape and textural features," *Sustainability*, vol. 9, no. 8, p. 1335, Aug. 2017.

[9] Y. Li, Z. Guo, F. Shuang, M. Zhang, and X. Li, "Key technologies of machine vision for weeding robots: A review and benchmark," *Comput. Electron. Agricult.*, vol. 196, May 2022, Art. no. 106880.

[10] Y. J. Wu, J. Wang, Y. L. Wang, Y. W. Zhao, and S. Zhang, "Field crop extraction based on machine vision," in *Proc. IEEE Int. Conf. Mechatronics Autom.*, Aug. 2021, pp. 1–5.

[11] A. Wang, W. Zhang, and X. Wei, "A review on weed detection using ground-based machine vision and image processing techniques," *Comput. Electron. Agricult.*, vol. 158, pp. 226–240, Mar. 2019.

[12] B. Niu, Q. Feng, B. Chen, C. Ou, Y. Liu, and J. Yang, "HSI-TransUNet: A transformer based semantic segmentation model for crop mapping from UAV hyperspectral imagery," *Comput. Electron. Agricult.*, vol. 201, Oct. 2022, Art. no. 107297.

[13] D. Stroppiana, P. Villa, G. Sona, G. Ronchetti, G. Candiani, M. Pepe, L. Busetto, M. Migliazzi, and M. Boschetti, "Early season weed mapping in rice crops using multi-spectral UAV data," *Int. J. Remote Sens.*, vol. 39, nos. 15–16, pp. 5432–5452, Aug. 2018.

[14] I. Sa, M. Popovic, R. Khanna, Z. Chen, P. Lottes, F. Liebisch, J. Nieto, C. Stachniss, A. Walter, and R. Siegwart, "WeedMap: A large-scale semantic weed mapping framework using aerial multispectral imaging and deep neural network for precision farming," *Remote Sens.*, vol. 10, no. 9, p. 1423, Sep. 2018.

[15] H. K. Suh, J. W. Hofstee, and E. J. van Henten, "Improved vegetation segmentation with ground shadow removal using an HDR camera," *Precis. Agricult.*, vol. 19, no. 2, pp. 218–237, Apr. 2018.

[16] Y. Lan, K. Huang, C. Yang, L. Lei, J. Ye, J. Zhang, W. Zeng, Y. Zhang, and J. Deng, "Real-time identification of rice weeds by UAV low-altitude remote sensing based on improved semantic segmentation model," *Remote Sens.*, vol. 13, no. 21, p. 4370, Oct. 2021.

[17] C. Nong, X. Fan, and J. Wang, "Semi-supervised learning for weed and crop segmentation using UAV imagery," *Frontiers Plant Sci.*, vol. 13, Jul. 2022, Art. no. 927368.

[18] K. Zou, X. Chen, Y. Wang, C. Zhang, and F. Zhang, "A modified U-Net with a specific data argumentation method for semantic segmentation of weed images in the field," *Comput. Electron. Agricult.*, vol. 187, Aug. 2021, Art. no. 106242.

[19] K. Zou, L. Ge, C. Zhang, T. Yuan, and W. Li, "Broccoli seedling segmentation based on support vector machine combined with color texture features," *IEEE Access*, vol. 7, pp. 168565–168574, 2019.

[20] M. R. Golzarian and R. A. Frick, "Classification of images of wheat, ryegrass and brome grass species at early growth stages using principal component analysis," *Plant Methods*, vol. 7, no. 1, p. 28, 2011.

[21] F. Kitzler, N. Barta, R. W. Neugschwandtner, A. Gronauer, and V. Motsch, "WE3DS: An RGB-D image dataset for semantic segmentation in agriculture," *Sensors*, vol. 23, no. 5, p. 2713, Mar. 2023.

[22] A. Abdalla, H. Cen, L. Wan, R. Rashid, H. Weng, W. Zhou, and Y. He, "Fine-tuning convolutional neural network with transfer learning for semantic segmentation of ground-level oilseed rape images in a field with high weed pressure," *Comput. Electron. Agricult.*, vol. 167, Dec. 2019, Art. no. 105091.

[23] S. Rehman, M. A. Khan, M. Alhaisoni, A. Armghan, F. Alenezi, A. Alqahtani, K. Vesal, and Y. Nam, "Fruit leaf diseases classification: A hierarchical deep learning framework," *Comput., Mater. Continua*, vol. 75, no. 1, pp. 1179–1194, 2023.

[24] S. Vijh, P. Gaurav, S. Kumar, P. Bansal, M. Singh, M. A. Khan, and V. Palade, "USMA-BOF: A novel bag-of-features algorithm for classification of infected plant leaf images in precision agriculture," *IEEE Robot. Autom. Mag.*, vol. 30, no. 4, pp. 30–40, Dec. 2023.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[26] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Computer-Assist. Intervent.*, Munich, Germany, vol. 9351, 2015, pp. 234–241.

[28] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[29] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[30] A. Kamilaris and F. X. Prenafeta-Boldu, "A review of the use of convolutional neural networks in agriculture," *J. Agricult. Sci.*, vol. 156, no. 3, pp. 312–322, Apr. 2018.

[31] P. Chen, R. Wang, and P. Yang, "Editorial: Deep learning in crop diseases and insect pests," *Frontiers Plant Sci.*, vol. 14, Feb. 2023, Art. no. 1145458.

[32] D. Liu, H. Yang, Y. Gong, and Q. Chen, "A recognition method of crop diseases and insect pests based on transfer learning and convolution neural network," *Math. Problems Eng.*, vol. 2022, pp. 1–10, Jul. 2022.

[33] J. Liu, M. R. Zhao, and X. F. Guo, "A fruit detection algorithm based on R-FCN in natural scene," in *Proc. 32nd Chin. Control Decis. Conf.*, 2020, pp. 487–492.

[34] K. Zou, Q. Liao, F. Zhang, X. Che, and C. Zhang, "A segmentation network for smart weed management in wheat fields," *Comput. Electron. Agricult.*, vol. 202, Nov. 2022, Art. no. 107303.

[35] M. Das and A. Bais, "DeepVeg: Deep learning model for segmentation of weed, canola, and canola flea beetle damage," *IEEE Access*, vol. 9, pp. 119367–119380, 2021.

[36] S. Jin, H. Dai, J. Peng, Y. He, M. Zhu, W. Yu, and Q. Li, "An improved mask R-CNN method for weed segmentation," in *Proc. IEEE 17th Conf. Ind. Electron. Appl. (ICIEA)*, Dec. 2022, pp. 1430–1435.

[37] K. Jiang, U. Afzaal, and J. Lee, "Transformer-based weed segmentation for grass management," *Sensors*, vol. 23, no. 1, p. 65, Dec. 2022.

[38] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, vol. 12346, 2020, pp. 213–229.

[39] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12159–12168.

[40] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[41] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Conf. Workshop Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12077–12090.

[42] Y. H. Kim and K. R. Park, "MTS-CNN: Multi-task semantic segmentation-convolutional neural network for detecting crops and weeds," *Comput. Electron. Agricult.*, vol. 199, Aug. 2022, Art. no. 107146.

[43] A. Nasiri, M. Omid, A. Taheri-Garavand, and A. Jafari, "Deep learning-based precision agriculture through weed recognition in sugar beet fields," *Sustain. Comput., Informat. Syst.*, vol. 35, Sep. 2022, Art. no. 100759.

[44] Y. Hao, Y. Liu, Y. Chen, L. Han, J. Peng, S. Tang, G. Chen, Z. Wu, Z. Chen, and B. Lai, "EISeg: An efficient interactive segmentation tool based on PaddlePaddle," 2022, *arXiv:2210.08788*.

[45] J. Gu, H. Kwon, D. Wang, W. Ye, M. Li, Y. H. Chen, L. Lai, V. Chandra, and D. Z. Pan, "Multi-scale high-resolution vision transformer for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 12084–12093.

[46] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11966–11976.

[47] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "ConvNeXt v2: Co-designing and scaling ConvNets with masked autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 16133–16142.

[48] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[49] S. H. Woo, J. Park, and J. Y. Lee, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 11211, 2018, pp. 3–19.

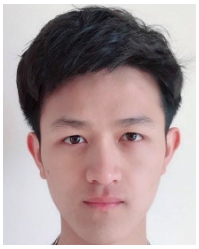
[50] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[51] D. Zhou, M. Li, Y. Li, J. Qi, K. Liu, X. Cong, and X. Tian, "Detection of ground straw coverage under conservation tillage based on deep learning," *Comput. Electron. Agricult.*, vol. 172, May 2020, Art. no. 105369.

[52] N. Chebrolu, P. Lottes, A. Schaefer, W. Winterhalter, W. Burgard, and C. Stachniss, "Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields," *Int. J. Robot. Res.*, vol. 36, no. 10, pp. 1045–1052, Sep. 2017.

CUIMIN SUN received the Ph.D. degree in computer science from the University of Science and Technology of China. She is currently a Lecturer with Guangxi University, specializing in image processing and pattern recognition. Her research interests include developing innovative algorithms and techniques for image analysis, feature extraction, and pattern recognition. Her work aims to improve the accuracy and efficiency of image processing systems, enabling applications in various domains, such as medical imaging, surveillance, and computer vision. Throughout her academic journey, she has published several research articles in reputable peer-reviewed journals and she has presented her findings at national and international conferences. Her contributions have been well-received and recognized by the scientific community.

MENGHUA ZHANG is currently pursuing the master's degree in computer science and technology with Guangxi University. He has five years of experience in software development. His research interests include computer vision and image processing.



MUCHEN ZHOU is currently pursuing the Graduate degree with the College of Mechanical Engineering, Guangxi University. He is familiar with mechanical equipment control and PLC programming technology. His research interests include image processing and machine learning.

XINGZHI ZHOU is currently pursuing the master's degree with Guangxi University. He specializes in the field of image processing and pattern recognition and is actively engaged in research within these domains.