IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 13 November 2023, accepted 9 December 2023, date of publication 18 December 2023, date of current version 22 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3344520

RESEARCH ARTICLE

Method for Segmentation of Bean Crop and Weeds Based on Improved UperNet

MINGYANG QI[1], HAOZHANG GAO[1,2], TETE WANG[3], BAOXIA DU[1,2], (Graduate Student Member, IEEE), HAN LI[1,2], WENYU ZHONG[1], AND YOU TANG[1]

[1]Electrical and Information Engineering College, Jilin Agricultural Science and Technology University, Jilin 132101, China
[2]School of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin 132022, China
[3]Research and Development Department, Jilin Province Electric Innovation Information Technology Company Ltd., Changchun 130117, China

Corresponding authors: Wenyu Zhong (zhongwenyu@jlnku.edu.cn) and You Tang (tangyou9000@163.com)

This work was supported in part by the Jilin Province Science and Technology Development Program Project under Grant YDZJ202201ZYTS692, and in part by the Digital Agriculture: An Emerging Interdisciplinary Field in Jilin Province under Grant 2018730519.

ABSTRACT Bean crop and weed detection is a key component of precision agriculture, which can distinguish between bean crop and weeds. Accurate identification of weeds is essential for precision weed management. This study introduces PF-UperNet, a semantic segmentation approach rooted in an encoder-decoder architecture, designed to autonomously distinguish between bean crop and weeds using advanced computer vision techniques. Our methodology refines the foundational UperNet in several significant ways: Firstly, we adopt the PoolFormer-S12 as a substitute for UperNet's backbone structure, aiming to reduce the model parameters and boost its performance metrics. Secondly, the Efficient Channel Attention (ECA) mechanism is integrated into both the PoolFormer-S12 and the Decoder, sharpening the network's focus on extracting salient channel features. Then, within the Decoder, the Feature Alignment Pyramid Network (FaPN) supplants the conventional Feature Pyramid Network (FPN) module, remedying the misalignment issues observed in UperNet's FPN feature maps. Lastly, we replace the Cross-Entropy loss with a combination of Cross-Entropy loss and Dice coefficient loss to increase the model's attention on regions to be detected. Empirical evidence underlines the efficacy of our technique, with a Mean Intersection over Union (MIoU) of 87.45%, a Mean Pixel Accuracy (MPA) of 96.82%, and a total of 46.16M parameters encapsulated. Relative to the benchmark UperNet, our model demonstrates enhancements of 1.08% and 0.25% in MIoU and MPA, respectively, and accomplishes a parameter reduction of 27.92%. Experimental results demonstrate that the proposed model achieves remarkable detection performance in terms of MIoU, MPA, and model parameters. It can provide an effective detection method for weed management.

INDEX TERMS Improved UperNet, green bean, weeds, semantic segmentation, PoolFormer, deep learning.

I. INTRODUCTION

The growth environment of crops is complex, and weeds often accompany the entire growth process of crops. During this period, weeds compete with crops for sunlight, water, nutrients, etc., severely affecting the yield and quality of crops [1]. Affected by pests, diseases, and weeds, global crop yields suffer nearly a 30% loss annually. Therefore, effective identification and removal of weeds are crucial for crop yield and quality [2], [3], [4], [5]. Traditional chemical weeding methods generally lack specificity, using overall spraying. This type of weeding adversely affects both the environment and the crops themselves [6], [7]. If chemical herbicides can be sprayed based on the distribution of weeds, it would not only reduce the use of chemical herbicides but also decrease pesticide residues in bean crop. The core issue of targeted chemical herbicide spraying is how to accurately identify bean crop and weeds. Therefore, researching the

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

segmentation tasks of bean crop and weeds is of paramount importance.

Until now, there are few studies on the segmentation of green beans and weeds. Yet, useful references and experiences could be obtained from the segmentation methods of other crops or other legumes and weeds. Multiple methods are available for the identification of crops and weeds, such as spectral and machine learning identification methods. However, due to the expensive equipment and complexity of spectral detection, it is not conducive to widespread adoption [8], [9]. Recently, as artificial intelligence technology has advanced, machine learning has become increasingly prevalent in crops and weeds identification tasks [10]. Machine learning classifies, detects, and segments the objects to be detected by describing data features and extracting useful information from them [11]. Common machine learning methods can be categorized into supervised and unsupervised learning, such as supervised learning's k-nearest neighbor (KNN) and logistic regression, and unsupervised learning's clustering and principal component analysis (PCA), etc. Islam et al. [12] employed aerial imagery from drones over chili pepper fields to identify weeds and evaluated the efficacy of algorithms like k-nearest neighbors (KNN), support vector machine (SVM), and random forest (RF). Bakhshipour and Jafari [13] introduced a technique for detecting weeds in beet fields. They integrated several morphological features, established a model for each plant type, and finally identified beets and weeds successfully using artificial neural networks(ANN) and SVM. While machine learning techniques are applicable for weeds identification, the intricate growth conditions of crops can be influenced by elements like lighting and climate, resulting in reduced detection precision. Moreover, machine learning requires extensive domain knowledge to construct features from data for detection [14]. Due to the reasons mentioned above, it is challenging to apply machine learning techniques, and the recognition accuracy is not high.

In recent years, the surge of deep convolutional neural networks (ConvNets) in computer vision has led to their widespread use in crop semantic segmentation tasks. Yang et al. [15] developed a multi-scale convolutional neural attention network named MSFCA-Net, which has been successfully applied to weeds identification in soybeans, beets, carrots, and rice. On their proprietary and public datasets, they achieved MIoUs of 92.64%, 89.58%, 79.34%, and 78.12%, respectively. Kamath et al. [16] proposed an improved PSPNet [17] and compared it with SegNet [18] and U-Net [19], successfully segmenting rice and weeds. Promising results have been achieved with an accuracy exceeding 90%. Targeted weed management was realized, reducing the harmful impact of herbicides on the environment. Zou et al. [20] introduced an enhanced U-Net semantic segmentation framework by adapting the backbone and reconfiguring the decoder architecture. They achieved segmentation of wheat and weeds with an IoU of 88.98% for the weeds, and the average detection speed on the edge device reached 52 FPS. Yu et al. [21] proposed an improved Deeplab V3+ model, successfully achieving segmentation of soybeans and weeds, with an MIoU of 91.53% on their custom dataset. Xu et al. [22] developed a segmentation network based on the encoder-decoder structure, integrating color indices with instance segmentation. It successfully segmented soybeans and weeds, achieving an MIoU of 93.9% on their custom dataset. Although ConvNets have achieved outstanding results in the segmentation tasks of crops and weeds, the receptive field of ConvNets is relatively small, resulting in room for improvement in the segmentation accuracy of crops and weeds. While some methods, such as pyramids [17], atrous convolutions [23], and attention mechanisms [24], have been proposed in ConvNets to compensate for the inadequacy of their receptive fields, researchers still hope for ConvNets to have a broader receptive field to achieve more superior detection performance.

Dosovitskiy et al. [25] introduced the Vision Transformer (ViT) to address the limited receptive field issue of ConvNets. Jiang et al. [26] replaced the traditional ConvNets with ViT to achieve weeds detection. Zhang et al. [27] utilized the improved version of ViT called Swin-Transformer [28] to construct an enhanced Swin-UNet segmentation model, successfully segmenting maize and weeds. At present, the ViT has achieved commendable segmentation results in tasks involving crops and weeds. However, its model parameters are large, consuming significant memory, which hinders model training and broader application. Furthermore, there is still room for improvement in model accuracy.

This study focuses on the identification of green bean and weeds. To ensure the detection speed and accuracy of green bean and weeds, a segmentation model with smaller parameters based on the ViT is required. Therefore, this study proposes an efficient semantic segmentation model for green bean and weeds, named PoolFormer-UperNet (PF-UperNet). Major adjustments and contributions of the model are as follows:

1) The backbone of UperNet [29] is replaced with PoolFormer-S12 [30], which not only reduces the model parameters but also expands the model's receptive field.

2) The ECA [31] attention module is added to the Encoder and Decoder, re-integrating the importance of channel information and enhancing the model's detection performance.

3) In UperNet, Feature Pyramid Network (FPN) [32] is replaced by Feature-aligned Pyramid Network (FaPN) [33], addressing the adverse effects caused by misalignment of feature maps during feature fusion.

4) Cross-Entropy loss is replaced with Cross-Entropy loss + Dice coefficient loss, making the segmentation model more focused on the areas to be detected.

To summarize in all, This model solves the problems that spectral detection is not conducive to promotion due to the high cost and complex operation of equipment, and machine learning requires a lot of professional domain knowledge to

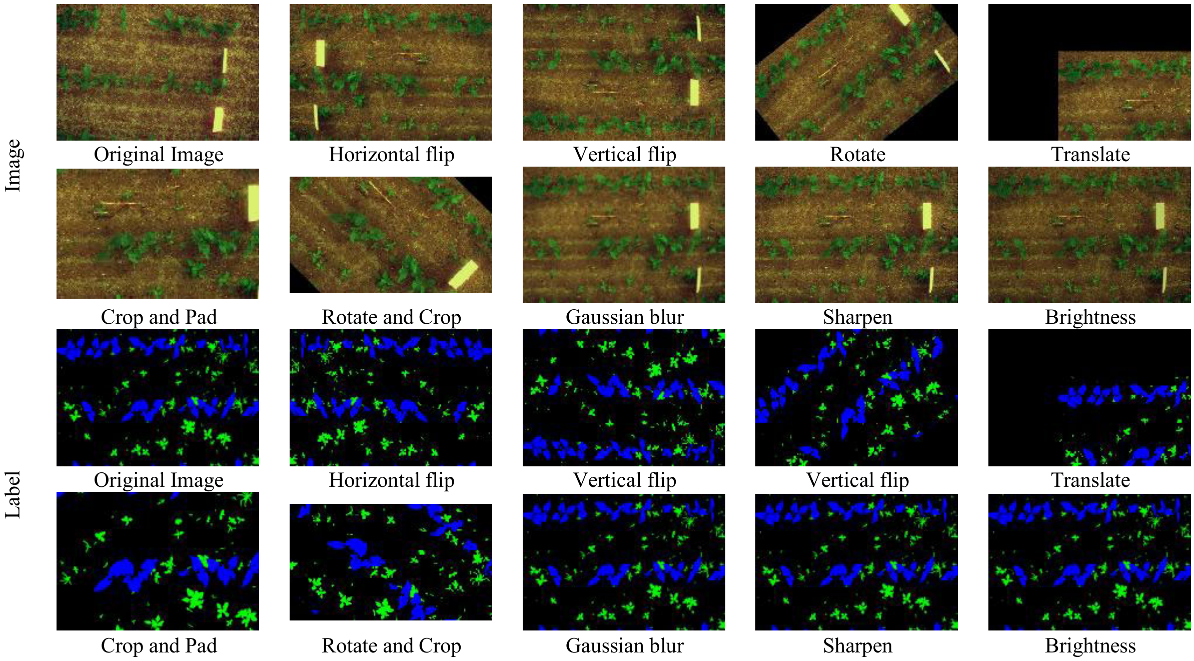


FIGURE 1. **Augmented dataset and its corresponding labels.**

build the features to be detected from the data. It also solves the problems that ConvNets have poor receptive field, ViT models have a large number of parameters and require a lot of GPU memory for training.

II. MATERIALS AND METHODS

A. DATASET

The dataset used in this paper is a public dataset, which was collected by Jehan-Antoine Vayssade using a multispectral camera in Airel, France (latitude 46°20’30.3”N, longitude 3°26’33.6”E). The multispectral camera is configured with bands at 450/570/675/710/730/850 nm, with an FWHM of 10 nm. The images showcase green bean alongside various native weeds such as yarrows, amaranth, geranium, plantago, etc. The collection conditions include rainy days, cloudy days, various lighting conditions, and different time periods. The aforementioned public dataset can be accessed at https://doi.org/10.15454/JMKP9S. The dataset contains a total of 300 folders, each of which contains one original image and one spectral image collected at 570nm, named as “false.png” and “image.tiff” respectively. There are two label files corresponding to the original images, named as “gt.png” and “gt.xml” respectively. In addition, there is one black-and-white image of green beans and weeds (not available in some folders) named as “index.png”. A total of 1376 files are contained in all the folders. After manually sorting the required “false.png” and “gt.png” for this study, we obtained a total of 300 pairs of bean crop and weed images and their corresponding label images. The above 300 pairs of images were divided into training set, validation set, and test set in the ratio of 7:1:2.

B. DATA ENHANCEMENT

To ensure better training performance and superior testing results with limited data, we adopted data augmentation techniques to expand the original dataset [34]. This paper utilized nine offline data augmentation methods, including horizontal flip, vertical flip, rotate, translate, crop and pad, rotate and crop, Gaussian blur, sharpen, and brightness. The original data, augmented dataset, and corresponding labels are shown in Figure 1.

C. OVERRALL STRUCTURE OF THE SEMANTIC SEGMENTATION MODEL

PoolFormer [30] is a general-purpose backbone for computer vision feature extraction, and its main principle is similar to that of Transformer. Yu et al. believe that the success of Transformer mainly stems from its overall structure. Therefore, they replaced self-attention with pooling, thereby maintaining excellent detection performance while reducing computational costs. Based on the outstanding performance of PoolFormer, this paper adopts PoolFormer as the backbone. The model is primarily divided into encoder and decoder sections. The specific structure is illustrated in Figure 2.

The input image is first processed by the PoolFormer structure. PoolFormer consists of 4 stages, each of which contains patch embedding, PoolFormer block, and ECA block. The ECA block is a channel attention module added after PoolFormer block to enhance the feature extraction ability of the backbone.

The features extracted by PoolFormer are passed into the Pyramid Pooling Module (PPM) and FaPN. In order

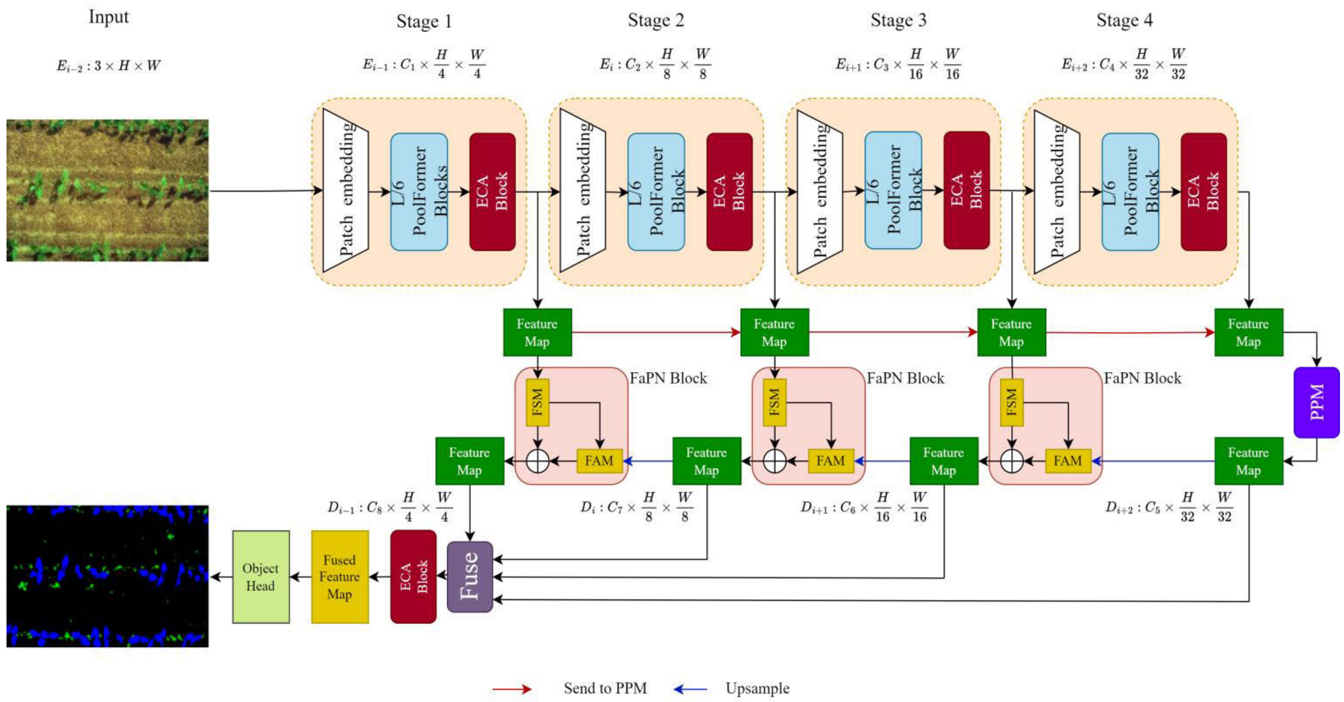

Input
$E_{i-2}: 3 \times H \times W$
Stage 1
$E_{i-1}: C_1 \times \frac{H}{4} \times \frac{W}{4}$
Stage 2
$E_i: C_2 \times \frac{H}{8} \times \frac{W}{8}$
Stage 3
$E_{i+1}: C_3 \times \frac{H}{16} \times \frac{W}{16}$
Stage 4
$E_{i+2}: C_4 \times \frac{H}{32} \times \frac{W}{32}$
Patch embedding
L/6 PoolFormer Blocks
ECA Block
L/6 PoolFormer Block
Feature Map
FaPN Block
FSM
FAM
PPM
$D_{i-1}: C_8 \times \frac{H}{4} \times \frac{W}{4}$
$D_i: C_7 \times \frac{H}{8} \times \frac{W}{8}$
$D_{i+1}: C_6 \times \frac{H}{16} \times \frac{W}{16}$
$D_{i+2}: C_5 \times \frac{H}{32} \times \frac{W}{32}$
Object Head
Fused Feature Map
Fuse
Send to PPM
Upsample

FIGURE 2. Overall structure of the semantic segmentation model.

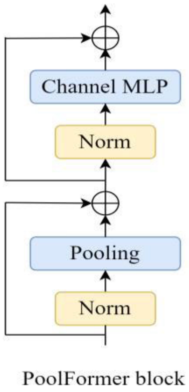


FIGURE 3. PoolFormer structure.

to augment the predictive proficiency of UperNet, we substituted its Feature Pyramid Network (FPN) with a Feature-aligned Pyramid Network (FaPN), thereby mitigating misalignment issues encountered during the amalgamation of backbone-derived features and feature maps. After processing through PPM and FaPN, the feature maps will be integrated via the fuse module, then connected to the ECA, and finally passed through the object head to produce the ultimate prediction results.

D. ENCODER

1) BACKBONE NETWORK

The backbone network consists of the stages 1-4 shown in Figure 2. The input image is reduced to half its original height and width after each stage, and the number of channels is increased to 64, 128, 320, and 512, respectively. Patch embedding is composed of a convolutional layer and a flattened layer. Each image patch enters the core module of PoolFormer. This module consists of the following steps: 1. Normalize the image patch using group normalization. 2. Pool the image patch using a pooling operation. 3. Pool the image patch again using a pooling operation. The result after the residual connection is again passed through group normalization and channel MLP. Then, the result is passed through the residual connection with the input from the previous level and output. The overall structure of PoolFormer is similar to that of Transformer, but it replaces self-attention with pooling operations. This design greatly reduces the computational cost. The structure diagram of PoolFormer is shown in Figure 3.

2) ECA BLOCK

The Efficient Channel Attention (ECA) block autonomously evaluates and assigns weights to channels within the acquired feature map, sequentially ranking them based upon their respective importances. This ensures that the model pays more attention to useful feature channels and suppresses unhelpful ones, enhancing segmentation performance. The operation process is as follows: Firstly, the input undergoes a Global Average Pooling (GAP) operation, resulting in a feature map with a height (H) and width (W) of 1. Then, 1×1 convolution is applied to the previously obtained feature map, effectively integrating information between channels.

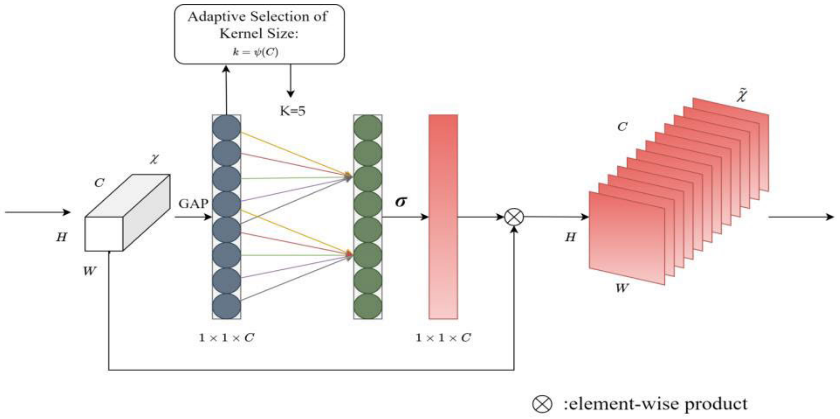


FIGURE 4. ECA block.

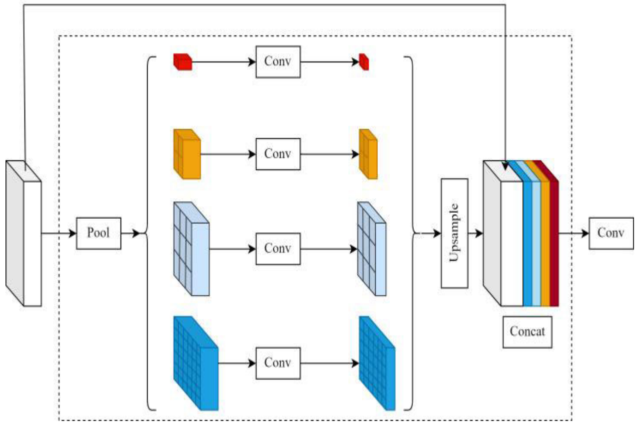


FIGURE 5. Pyramid pooling module.

Finally, the results from the 1×1 convolution are passed through a sigmoid function and then multiplied with the input. As only 1×1 convolution operation was used without introducing other complex convolutions, the computational cost of the ECA block is relatively low, which can enhance the model's accuracy while having minimal impact on model parameters. The principle of the ECA block used in this paper is shown in Figure 4.

E. DECODER

1) PPM BLOCK

PoolFormer has expanded the model's receptive field to some extent, while we hope the model can achieve better segmentation results at different scales. The pyramid pooling module (PPM) fully integrates contextual information, allowing objects of different scales to achieve better segmentation performance. The operation process is as follows: Firstly, the feature map is divided into different image blocks, including 1×1, 2×2, 3×3, and 6×6. Then, different pooling operations are applied to each image block. Next, the number of channels is adjusted through 1×1 convolution, followed by upsampling. Lastly, various feature maps are merged with the primary map, yielding a feature representation rich in contextual details. Figure 5 illustrates the operational mechanism of PPM.

2) FaPN BLOCK

In the original UperNet, the Feature Pyramid Network (FPN) is utilized. However, in FPN, directly adding the upsampled features with those from the backbone can result in misaligned contextual feature maps, leading to segmentation classification errors, especially evident at the edges. Hence, this study introduces the Feature Alignment Pyramid Network (FaPN). By aligning the contextual features through FaPN before summation, we effectively address the misalignment issue. FaPN comprises two modules: Feature Selector Module (FSM) and Feature Alignment Module (FAM). Their working principles are shown in Figures 6 and 7, respectively. Specifically, the FSM process involves the following steps: First, E_i extracts the significance of each channel through f_m in the FSM module and multiplies it with E_i. Next, the result from the previous step is added to E_i. Lastly, channel adjustments are made through f_s(1×1 convolution), resulting in $\hat{E}_i$.

The input to the FAM model consists of two parts: one is derived from the output $\hat{E}_i$ of the FSM model, and the other comes from the prior-level feature map shown in Figure 2. After upsampling, we refer to this part as D_i. The operational process is as follows: First, D_i, after being upsampled to D_i^u, is concatenated with $\hat{E}_i$, followed by a convolution operation, resulting in an offset-inclusive image Δ_i. Then, Δ_i and D_i^u are jointly fed into the deformable convolution for feature alignment, yielding the output $\hat{D}_i^u$. Lastly, pixels of $\hat{D}_i^u$ and $\hat{E}_i$ are added together, completing the entire FaPN process. FaPN, through continuous training, enables convolution to detect misaligned portions between two images and generate an offset. This offset is then used for alignment, thereby compensating for the deficiencies of FPN.

Moreover, the fuse block in Figure 2 integrates the outputs of PPM and FaPN into a consistent image size using

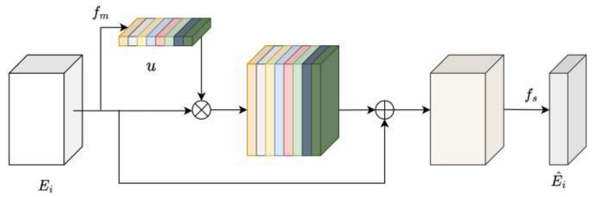


FIGURE 6. **Feature selector module.**

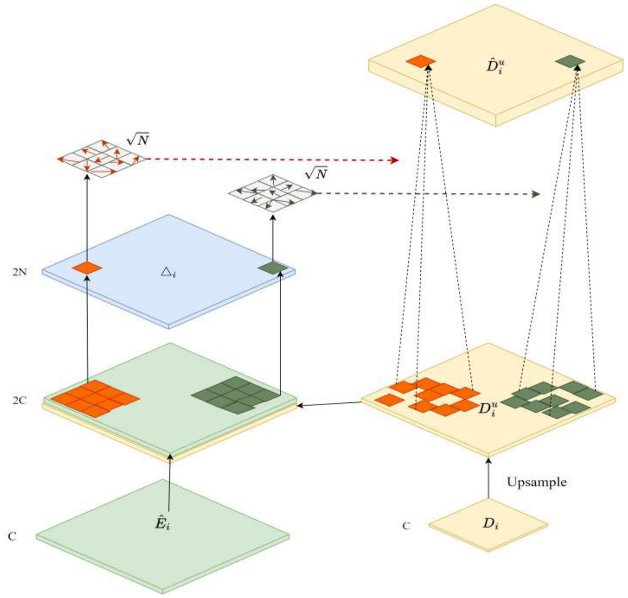


FIGURE 7. **Feature alignment module.**

bilinear interpolation, subsequently passing through the ECA block for feature channel importance adjustment. Ultimately, by convolutional operations, features from different stages and modules are integrated, producing the prediction results of PF-UperNet.

F. TRANSFER LEARNING

Transfer learning is a technique in machine learning where a model designed for one task (task A) is repurposed as a foundation for a different task (task B). The main idea is to use strategies from previously solved problems to address unsolved ones. Due to the small size of the green bean and weeds dataset in this study, to improve the training results, therefore, transfer learning was employed in the training process of this study. Specifically, we pre-trained the PoolFormer-S12 on the ADE20K dataset. This pre-training allowed us to transfer the knowledge gained from ADE20K to our green bean and weeds dataset, enhancing the feature extraction process for these plants.

G. LOSS FUNCTION

In this paper, the loss function is denoted as L_T, comprising two components: L_1 and L_2. L_1 represents the c, and L_2 represents the Dice coefficient loss. Additionally, w_1 and w_2 are used to weight L_1 and L_2 respectively. In this paper, $w_1 = 0.6$ and $w_2 = 0.4$. The loss function is computed using the following equation:

$$L_1 = -\frac{1}{N}\sum\nolimits_{i=1}^{N}\sum\nolimits_{j=1}^{M} y_{ic}\ln(p_{ic}) \tag{1}$$

$$L_2 = 1 - \frac{2|X \cap Y| + \varepsilon}{|X| + |Y| + \varepsilon} \tag{2}$$

$$L_T = w_1 L_1 + w_2 L_2 \tag{3}$$

In expression (L_1), the variables are defined as follows: N is the batch size, M is the total number of categories, and y_{ic} is a binary sign function, taking a value of 1 when the true category of sample i matches c, and 0 otherwise. The term p_{ic} refers to the predicted probability that sample i is assigned to category c.

In expression (L_2), X denotes the ground truth while Y denotes the category predicted by the semantic segmentation algorithm. To prevent division by zero in the denominator, a safeguard value ε is employed, with ε set to 1×10^{-6}.

H. EVALUATION METRICS

In this study, we use Mean Pixel Accuracy (MPA), Intersection over Union (IoU), and Mean Intersection over Union (MIoU) as our evaluation criteria for semantic segmentation, outlined in equations (4) to (6) [35]. MPA quantifies the proportion of pixels accurately identified for each category to the total pixel count. IoU measures the overlap between the predicted and actual values for a given category, representing the proportion of shared area to their combined area. MIoU, on the other hand, is the average overlap across all categories, determined by summing the IoU values and then computing their mean. These benchmarks help gauge the efficacy of our model's segmentation.

$$MPA = \frac{1}{k+1}\sum\nolimits_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \times 100\% \tag{4}$$

$$IoU = \sum\nolimits_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \times 100\% \tag{5}$$

$$MIoU = \frac{1}{k+1}\sum\nolimits_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \times 100\% \tag{6}$$

In the aforementioned formula, p_{ii} represents the number of pixels that are actually of category I and are predicted as I, p_{ij} represents the number of pixels that are actually of category I but are predicted as J, p_{ji} represents the number of pixels that are actually of category J but are predicted as I, k represents the number of categories, and in this paper, $k = 2$.

III. RESULTS AND ANALYSIS

A. EXPERIMENTAL ENVIRONMENT AND TRAINING PARAMETER SETTINGS

To maintain the consistency and reliability of our experimental findings, every test was carried out under identical

TABLE 1. Ablation results of the UperNet.

Network	Model	PoolFormer	CE+Dice	FaPN	ECA	MIoU(%)	MPA(%)	Background IoU(%)	Green Bean IoU(%)	Weeds IoU(%)
	Model1	×	×	×	×	86.37	96.57	96.72	77.92	84.46
	Model2	√	×	×	×	87.17	96.79	96.93	79.18	85.4
	Model3	√	√	×	×	87.28	96.82	96.99	79.25	85.62
	Model4	√	√	√	×	87.36	96.83	96.99	79.48	85.51
UperNet	Model5	√	√	√	√	87.45	96.82	96.91	79.79	85.65

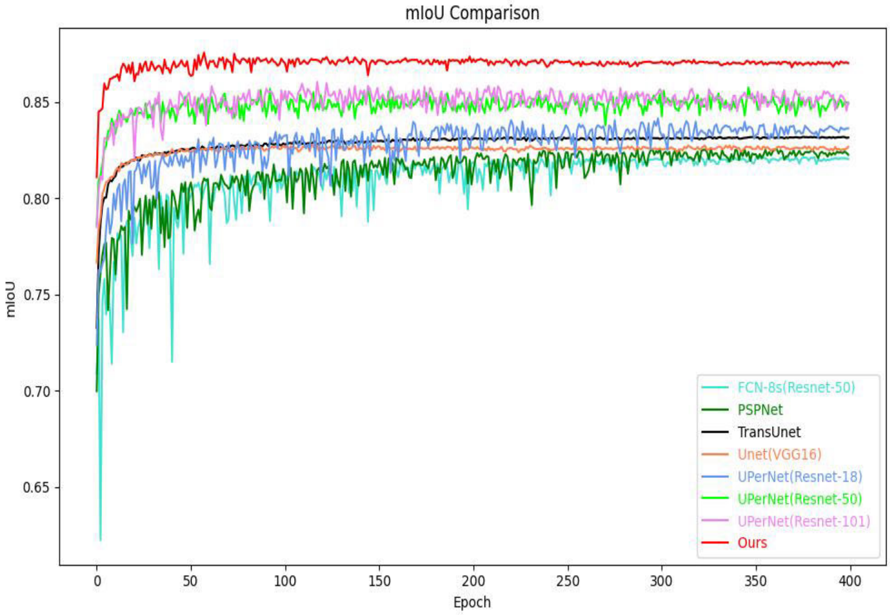


FIGURE 8. MIoU trend curve for different semantic segmentation models during validation.

conditions. Our setup operated on Ubuntu 22.04.2 LTS, powered by an Intel(R) Xeon(R) Silver 4110 CPU clocked at 2.10GHz, and utilized an NVIDIA GTX3090 GPU with 24GB of video memory. The programming language used was Python3.8.17, with PyTorch1.10.0 as the deep learning framework. In the experiments, the CUDA11.1 architecture was adopted as the unified computing device framework.

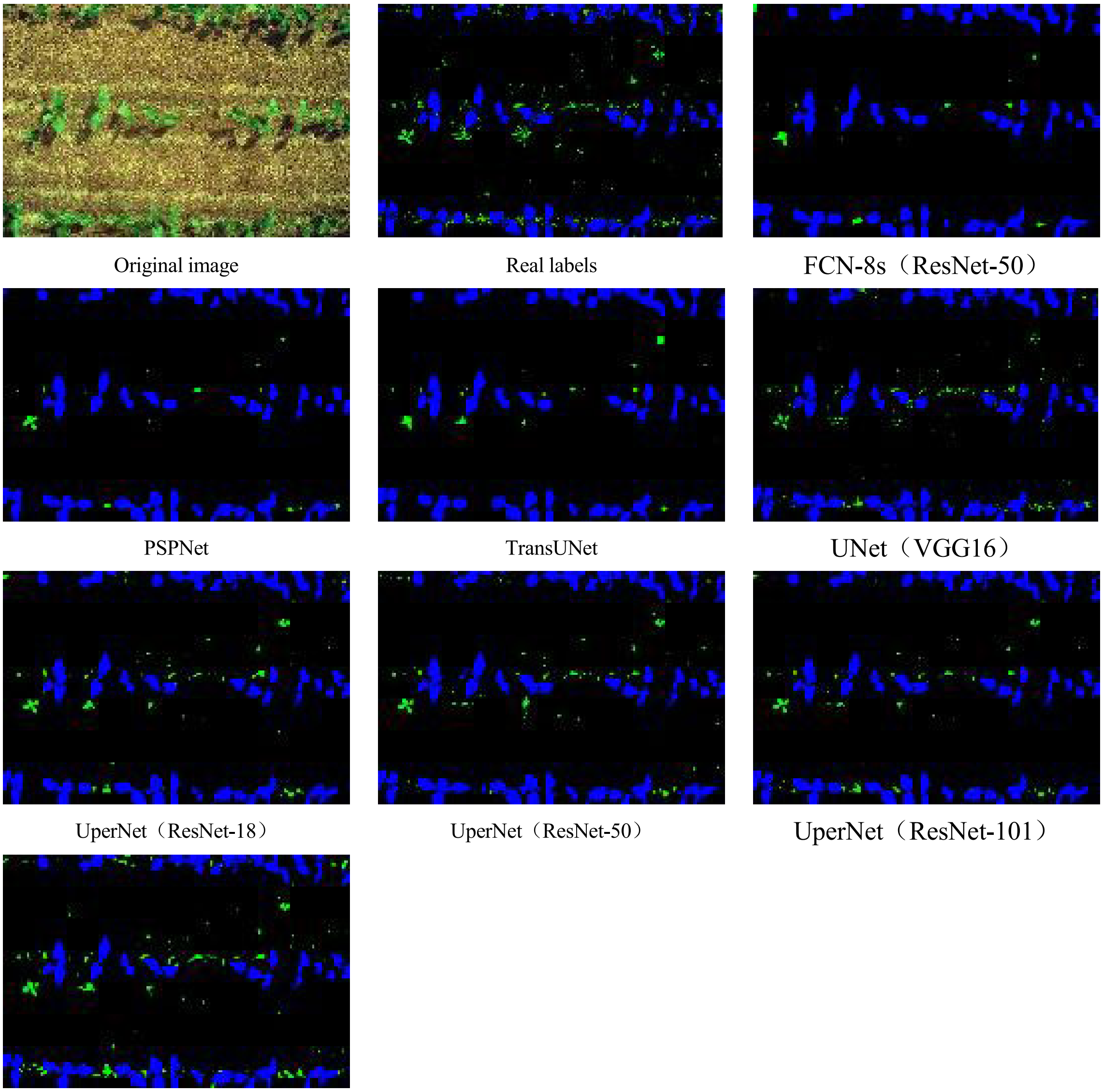
The training parameters for this study are as follows: Maximum training epochs was 400, Batch size was 12, Optimizer was Adaptive Moment Estimation with Weight Decay (AdamW), Initial learning rate was 0.0002, Learning rate weight decay coefficient was 0.0001.

B. ABLATION EXPERIMENT

In the original UperNet, ResNet50 [36] was used as the backbone with Cross-Entropy loss as the loss function. Notably, CE+Dice represents the simultaneous use of Cross-Entropy loss and Dice coefficient loss, with respective weights of 0.6 and 0.4. The performance of different modules was evaluated in the experiment, and specific results are shown in Table 1, where "√" in the table indicates the use of that module and "×" indicates the module was not used.

Model1 in Table 1 represents the classic UperNet, while model5 represents the final improved model PF-UperNet proposed in this paper. Observing Table 1, it is evident that the original UperNet achieved an MIoU of 86.37% and an MPA of 96.57% in the segmentation tasks of green bean and weeds. The PF-UperNet(model5) semantic segmentation model proposed in this paper achieved performances of 87.45% in MIoU and 96.82% in MPA respectively. Compared to Model1, PF-UperNet improved by 1.08% in MIoU and 0.25% in MPA. This enhancement effectively optimized the classic UperNet, enhancing its semantic segmentation performance.

Model2 initially replaces the classic UperNet (Model1)'s ResNet50 with PoolFormer, expanding the model's receptive field and reducing the model size. These improvements resulted in a 0.8% and 0.22% increase in MIoU and MPA for Model2 compared to Model1, respectively. Building on Model2, the Cross-Entropy loss is replaced with 0.6 Cross-Entropy loss+ 0.4 Dice coefficient loss to emphasize the region of interest. This model is termed Model3. Compared to Model2, Model3 has achieved a 0.11% increase in MIoU and a 0.03% increase in MPA. Model4, based on Model3,



Original image

Real labels

FCN-8s（ResNet-50）

PSPNet

TransUNet

UNet（VGG16）

UperNet（ResNet-18）

UperNet（ResNet-50）

UperNet（ResNet-101）

Ours

FIGURE 9. Test results of various semantic segmentation models.

introduces the FaPN block to address the feature misalignment issue caused by directly using FPN in UperNet. Relative to Model3, Model4 shows a 0.08% increase in MIoU and a 0.01% increase in MPA. Model5, building on Model4, incorporates the ECA block, re-integrating the importance information of the feature map channels. Although Model5 shows a slight decrease of 0.01% in MPA compared to Model4, its impact on segmentation results is negligible, while MIoU increases by 0.09%. The final optimized UperNet (Model5) achieved 87.45% and 96.82% in MIoU and MPA, respectively. These results indicate that, after a series of improvements, PF-UperNet has achieved significant performance enhancements in the semantic segmentation tasks of green bean and weeds.

C. COMPARISON WITH DIFFERENT NETWORK ARCHITENTURES

To further assess the effectiveness of our constructed model, we chose several commonly used semantic segmentation models for comparison with our PF-UperNet. These common semantic segmentation models include five in total: FCN [37], PSPNet [17], TransUNet [38], U-Net [19], and UperNet [29]. It should be noted that UperNet employed different backbones for experiments, thus we covered a total of 7 distinct semantic segmentation models. The MIoU performance of these models on the validation set during the training process is shown in Figure 8. From Figure 8, we can observe that the MIoU values of these models tend to stabilize as the training epochs increase, while the MIoU values of

TABLE 2. Results of different semantic segmentation models.

Model	Backbone Network	MIoU(%)	MPA(%)	Background IoU(%)	Green Bean IoU(%)	Weeds IoU(%)	Parameters/M
FCN-8s	ResNet50	81.92	95.25	95.26	70.49	80.02	47.11
U-Net	VGG16	83.07	95.89	96.42	71.33	81.46	29.06
PSPNet	ResNet50	82.5	95.37	95.28	71.61	80.61	46.58
TransUNet	Transformer	82.63	95.28	95.24	71.87	80.77	93.23
UperNet	ResNet18	84.33	96.08	96.35	74.06	82.59	40.78
UperNet	ResNet50	86.37	96.57	96.72	77.92	84.46	64.04
UperNet	ResNet101	86.2	96.54	96.64	77.51	84.46	83.03
Ours	Improved PoolFormer	87.45	96.82	96.91	79.79	85.65	46.16

PF-UperNet are consistently superior to other models across different epochs.

We further assessed our proposed model's performance using the test set. The evaluation results of different semantic segmentation models on the test set are shown in Table 2. As can be seen from Table 2, our proposed PF-UperNet outperforms other models in terms of MIoU, MPA, background IoU, green bean IoU, and weeds IoU evaluation metrics. To be precise, the MIoU of PF-UperNet is 87.45%, the MPA is 96.82%, the background IoU is 96.91%, the green bean IoU is 79.79%, and the weeds IoU is 85.65%. When juxtaposed with the least efficient FCN-8s model, PF-UperNet trims down model parameters by 0.95M and escalates MIoU and MPA by 5.53% and 1.57% respectively. When compared with U-Net, boasting the minimum model parameters, PF-UperNet, although augmenting parameter count by 17.1M, boosts MIoU and MPA by 4.38% and 0.93% respectively. In comparison to PSPNet, PF-UperNet scales down model parameters by 0.42M and sees a rise in MIoU and MPA by 4.95% and 1.45% respectively. When set against UperNet (ResNet18), PF-UperNet, albeit escalating model parameters by 5.38M, uplifts MIoU and MPA by 3.12% and 0.74% respectively. TransUNet, UperNet (ResNet50), and UperNet (ResNet101) each possess greater model parameters than PF-UperNet. Moreover, their metrics like MIoU, MPA, background IoU, green bean IoU, and weeds IoU also trail behind PF-UperNet. In conclusion, the PF-UperNet we proposed achieves satisfactory performance across all metrics.

To more intuitively demonstrate the improvements of our model, we have created test result images, which include the original images and true labels, as well as FCN-8s (ResNet-50), PSPNet, TransUNet, U-Net (VGG16), UperNet (ResNet-18), UperNet (ResNet-50), UperNet (ResNet-101), and our proposed model. These test results are shown in Figure 9.

IV. CONCLUSION AND DISSCUSION

This paper proposed an enhanced UperNet model for the segmentation of green bean and weeds. Firstly, we employed PoolFormer as a replacement for UperNet's backbone to reduce model parameters and enhance evaluation metrics. Secondly, we incorporated the ECA attention module in both PoolFormer and the decoder to amplify the model's focus on significant channel information. Subsequently, we substituted the FPN in UperNet with FaPN to address the misalignment of features during upsampling and the fusion of low-level feature maps. Lastly, we exchanged the Cross-Entropy loss for the combination of Cross-Entropy loss and Dice coefficient loss to achieve improved segmentation results for green bean and weeds.

This model addresses the problem that machine learning requires a lot of domain knowledge to construct the features to be detected from data. In addition, it also addresses the problems that ConvNets have small receptive fields, and ViT models have large parameter sizes and require a lot of GPU memory for training. According to the experimental results in this paper, the improved UperNet model's green bean and weeds segmentation method achieves an MIoU of 87.45 and MPA of 96.82. Additionally, it records background IoU, green bean IoU, and weeds IoU of 96.91%, 79.79%, and 85.65% respectively. The model has 46.16M parameters. This indicates that the improved UperNet model can effectively segment green bean and weeds, providing technical support for targeted weeding.

In this study, our model has achieved good performance on the Green Bean and Weeds dataset. However, the performance may degrade when applied to other bean crops and weeds. Additionally, our study requires a certain distance between the camera and the objects to ensure image quality. Therefore, it may not be suitable for applications in the drone field. In future research, we plan to create a dataset for green bean and weeds to enrich data resources and to thereby improve detection performance. To ensure applicability on edge devices, we also plan to further reduce model parameters while continuing to optimize model evaluation metrics, aiming to enhance real-time image segmentation performance. Furthermore, when combining Cross-Entropy loss and Dice coefficient loss, further research is needed to determine when their respective weights are optimal and how to select the optimal weight parameters.

APPENDIX

CODE

The code is available at https://github.com/MingyangQi1/PF-UperNet/tree/main

REFERENCES

[1] X. Zhang, J. Cui, H. Liu, Y. Han, H. Ai, C. Dong, J. Zhang, and Y. Chu, ''Weed identification in soybean seedling stage based on optimized faster R-CNN algorithm,'' *Agriculture*, vol. 13, no. 1, p. 16, Jan. 2023.

[2] S. Khan, M. Tufail, M. T. Khan, Z. A. Khan, and S. Anwar, ''Deep learning-based identification system of weeds and crops in strawberry and pea fields for a precision agriculture sprayer,'' *Precis. Agricult.*, vol. 22, no. 6, pp. 1711–1727, Dec. 2021.

[3] M. Dadashzadeh, Y. Abbaspour-Gilandeh, T. Mesri-Gundoshmian, S. Sabzi, J. L. Hernández-Hernández, M. Hernández-Hernández, and J. I. Arribas, ''Weed classification for site-specific weed management using an automated stereo computer-vision machine-learning system in rice fields,'' *Plants*, vol. 9, no. 5, p. 559, Apr. 2020.

[4] M. Alam, M. S. Alam, M. Roman, M. Tufail, M. U. Khan, and M. T. Khan, ''Real-time machine-learning based crop/weed detection and classification for variable-rate spraying in precision agriculture,'' in *Proc. 7th Int. Conf. Electr. Electron. Eng. (ICEEE)*, Apr. 2020, pp. 273–280.

[5] Y.-H. Tu, K. Johansen, S. Phinn, and A. Robson, ''Measuring canopy structure and condition using multi-spectral UAS imagery in a horticultural environment,'' *Remote Sens.*, vol. 11, no. 3, p. 269, Jan. 2019.

[6] A. Brilhador, M. Gutoski, L. T. Hattori, A. de Souza Inacio, A. E. Lazzaretti, and H. S. Lopes, ''Classification of weeds and crops at the pixel-level using convolutional neural networks and data augmentation,'' in *Proc. IEEE Latin Amer. Conf. Comput. Intell. (LA-CCI)*, Nov. 2019, pp. 93–98.

[7] Z. Wu, Y. Chen, B. Zhao, X. Kang, and Y. Ding, ''Review of weed detection methods based on computer vision,'' *Sensors*, vol. 21, no. 11, p. 23, May 2021.

[8] N. Sulaiman, N. N. Che'Ya, M. H. M. Roslim, A. S. Juraimi, N. M. Noor, and W. F. F. Ilahi, ''The application of hyperspectral remote sensing imagery (HRSI) for weed detection analysis in rice fields: A review,'' *Appl. Sci.*, vol. 12, no. 5, p. 2570, Mar. 2022.

[9] X. Zhao, X. Wang, C. Li, H. Fu, S. Yang, and C. Zhai, ''Cabbage and weed identification based on machine learning and target spraying system design,'' *Frontiers Plant Sci.*, vol. 13, Aug. 2022, Art. no. 924973.

[10] I. D. García-Santillán and G. Pajares, ''On-line crop/weed discrimination through the Mahalanobis distance from images in maize fields,'' *Biosyst. Eng.*, vol. 166, pp. 28–43, Feb. 2018.

[11] S. Bargoti and J. P. Underwood, ''Image segmentation for fruit detection and yield estimation in apple orchards,'' *J. Field Robot.*, vol. 34, no. 6, pp. 1039–1060, Sep. 2017.

[12] N. Islam, M. M. Rashid, S. Wibowo, C.-Y. Xu, A. Morshed, S. A. Wasimi, S. Moore, and S. M. Rahman, ''Early weed detection using image processing and machine learning techniques in an Australian chilli farm,'' *Agriculture*, vol. 11, no. 5, p. 13, Apr. 2021.

[13] A. Bakhshipour and A. Jafari, ''Evaluation of support vector machine and artificial neural networks in weed detection using shape features,'' *Comput. Electron. Agricult.*, vol. 145, pp. 153–160, Feb. 2018.

[14] A. S. M. M. Hasan, F. Sohel, D. Diepeveen, H. Laga, and M. G. K. Jones, ''A survey of deep learning techniques for weed detection from images,'' *Comput. Electron. Agricult.*, vol. 184, May 2021, Art. no. 106067.

[15] Q. Yang, Y. Ye, L. Gu, and Y. Wu, ''MSFCA-Net: A multi-scale feature convolutional attention network for segmenting crops and weeds in the field,'' *Agriculture*, vol. 13, no. 6, p. 1176, May 2023.

[16] R. Kamath, M. Balachandra, A. Vardhan, and U. Maheshwari, ''Classification of paddy crop and weeds using semantic segmentation,'' *Cogent Eng.*, vol. 9, no. 1, p. 18, Dec. 2022.

[17] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, ''Pyramid scene parsing network,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[18] 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.

[19] O. Ronneberger, P. Fischer, and T. Brox, ''U-net: Convolutional networks for biomedical image segmentation,'' in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Oct. 2015, pp. 234–241.

[20] 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.

[21] H. Yu, M. Che, H. Yu, and J. Zhang, ''Development of weed detection method in soybean fields utilizing improved DeepLabv3+ platform,'' *Agronomy*, vol. 12, no. 11, p. 15, Nov. 2022.

[22] B. Xu, J. Fan, J. Chao, N. Arsenijevic, R. Werle, and Z. Zhang, ''Instance segmentation method for weed detection using UAV imagery in soybean fields,'' *Comput. Electron. Agricult.*, vol. 211, Aug. 2023, Art. no. 107994.

[23] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, ''Semantic image segmentation with deep convolutional nets and fully connected CRFs,'' 2014, *arXiv:1412.7062*.

[24] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, ''CCNet: Criss-cross attention for semantic segmentation,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly, ''An image is worth 16×16 words: Transformers for image recognition at scale,'' 2020, *arXiv:2010.11929*.

[26] K. Jiang, U. Afzaal, and J. Lee, ''Transformer-based weed segmentation for grass management,'' *Sensors*, vol. 23, no. 1, p. 15, Jan. 2023.

[27] J. Zhang, J. Gong, Y. Zhang, K. Mostafa, and G. Yuan, ''Weed identification in maize fields based on improved Swin-Unet,'' *Agronomy*, vol. 13, no. 7, p. 15, Jul. 2023.

[28] 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.

[29] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, ''Unified perceptual parsing for scene understanding,'' in *Proc. Eur. Conf. Comp. Vis. (ECCV)*, Sep. 2018, pp. 418–434.

[30] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, ''MetaFormer is actually what you need for vision,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10809–10819.

[31] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, ''ECA-Net: Efficient channel attention for deep convolutional neural networks,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.

[32] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, ''Feature pyramid networks for object detection,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[33] S. Huang, Z. Lu, R. Cheng, and C. He, ''FaPN: Feature-aligned pyramid network for dense image prediction,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 844–853.

[34] J. He, J. Duan, Z. Yang, J. Ou, X. Ou, S. Yu, M. Xie, Y. Luo, H. Wang, and Q. Jiang, ''Method for segmentation of banana crown based on improved DeepLabv3+,'' *Agronomy*, vol. 13, no. 7, p. 1838, Jul. 2023.

[35] J. Sun, J. Zhou, Y. He, H. Jia, and Z. Liang, ''RL-DeepLabv3+: A lightweight rice lodging semantic segmentation model for unmanned rice harvester,'' *Comput. Electron. Agricult.*, vol. 209, Jun. 2023, Art. no. 107823.

[36] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[37] J. Long, E. Shelhamer, and T. Darrell, ''Fully convolutional networks for semantic segmentation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[38] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. Yuille, and Y. Zhou, ''TransUNet: Transformers make strong encoders for medical image segmentation,'' 2021, *arXiv:1412.7062*.

MINGYANG QI received the master's degree from the Changchun University of Technology, Changchun, China, in 2015. He is currently a Lecturer with the Electrical and Information Engineering College, Jilin Agricultural Science and Technology University, Jilin, China. His current research interests include deep learning and computer vision.

HAOZHANG GAO received the bachelor's degree from Shanghai Maritime University, Shanghai, China, in 2020. He is currently pursuing the M.S. degree with the School of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin, China. His current research interests include deep learning and computer vision.

TETE WANG is currently with the Research and Development Department, Jilin Province Electric Innovation Information Technology Company Ltd., Changchun, China. Her current research interest includes the analysis and application of agricultural big data.

BAOXIA DU (Graduate Student Member, IEEE) received the bachelor's degree from the Shandong University of Science and Technology, Qingdao, China, in 2020. He is currently pursuing the M.S. degree with the School of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin, China. His current research interests include deep learning, computer vision, and semantic communication.

HAN LI received the bachelor's degree from the Henan University of Engineering, Henan, China, in 2021. He is currently pursuing the M.S. degree with the School of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin, China. His current research interests include deep learning and computer vision.

WENYU ZHONG received the bachelor's degree from the Changchun University of Technology, Changchun, China, in 1998. His research interests include machine learning and agricultural automation.

YOU TANG received the Ph.D. degree from Northeast Agricultural University, Harbin, China, in 2017. He is currently a Professor with the Electrical and Information Engineering College, Jilin Agricultural Science and Technology University, Jilin, China. His research interests include bioinformatics and software engineering.