frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 27 March 2023
DOI 10.3389/fpls.2023.1123410

OPEN ACCESS

EDITED BY
Mehedi Masud,
Taif University, Saudi Arabia

REVIEWED BY
Jin Chen,
University of Kentucky, United States
Jana Shafi,
Prince Sattam Bin Abdulaziz University, Saudi Arabia

*CORRESPONDENCE
Huaxiang Liu
felicia_liu@126.com
Youyao Fu
fuyouyao828@126.com

SPECIALTY SECTION
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

RECEIVED 10 January 2023
ACCEPTED 03 March 2023
PUBLISHED 27 March 2023

CITATION
Fang J, Jiang H, Zhang S, Sun L, Hu X, Liu J, Gong M, Liu H and Fu Y (2023) BAF-Net: Bidirectional attention fusion network *via* CNN and transformers for the pepper leaf segmentation. *Front. Plant Sci.* 14:1123410. doi: 10.3389/fpls.2023.1123410

BAF-Net: Bidirectional attention fusion network *via* CNN and transformers for the pepper leaf segmentation

Jiangxiong Fang[1], Houtao Jiang[2], Shiqing Zhang[1], Lin Sun[2], Xudong Hu[2,3], Jun Liu[4], Meng Gong[2], Huaxiang Liu[1]* and Youyao Fu[1]*

[1]Institute of Intelligent Information Processing, Taizhou University, Taizhou, Zhejiang, China, [2]Jiangxi Engineering Laboratory on Radioactive Geoscience and Big Data Technology, East China University of Technology, Nanchang, China, [3]Engineering Research Center of Development and Management for Low to Ultra-Low Permeability Oil & Gas Reservoirs in West China, Xi'an Shiyou University, Xi'an, China, [4]College of Mechanical Engineering, Quzhou University, Quzhou, Zhejiang, China

The segmentation of pepper leaves from pepper images is of great significance for the accurate control of pepper leaf diseases. To address the issue, we propose a bidirectional attention fusion network combing the convolution neural network (CNN) and Swin Transformer, called BAF-Net, to segment the pepper leaf image. Specially, BAF-Net first uses a multi-scale fusion feature (MSFF) branch to extract the long-range dependencies by constructing the cascaded Swin Transformer-based and CNN-based block, which is based on the U-shape architecture. Then, it uses a full-scale feature fusion (FSFF) branch to enhance the boundary information and attain the detailed information. Finally, an adaptive bidirectional attention module is designed to bridge the relation of the MSFF and FSFF features. The results on four pepper leaf datasets demonstrated that our model obtains F1 scores of 96.75%, 91.10%, 97.34% and 94.42%, and IoU of 95.68%, 86.76%, 96.12% and 91.44%, respectively. Compared to the state-of-the-art models, the proposed model achieves better segmentation performance. The code will be available at the website: https://github.com/fangchj2002/BAF-Net.

KEYWORDS

convolution neural network, leaf segmentation, attention mechanism, multi-scale network, Swin Transformer

1 Introduction

Pepper is a common crop in China and has become an important vegetable and condiment in our daily life. However, pepper is a sensitive plant and pepper crops are highly exposed to diseases, which easily cause the frontal disease of the pepper leaves. The plant leaves can reflect plant growth, and pepper leaf diseases directly leads to the decline of

pepper yield and quality. The visual characteristics of pepper leaf diseases is very similar, so it is not easy to distinguish them. With the advance of imaging technology, computer vision technologies have been widely used in plant leaf extraction to guide the agricultural expert to analyze the crop growth. By using image processing technology to analyze two-dimensional leaf image features, the plant growth stages could be dissected (Slaughter et al., 2008; Koirala et al., 2019), and monitor the plant diseases (Singh, 2019; Tian et al., 2019) by the analysis of the image various plant organs. Therefore, the accurate segmentation of pepper leaves from pepper images is of great significance for controlling pepper leaf diseases. However, it is challenging to design a general model for automatic segmentation of pepper leaves since the pepper leaves and some crops have similar phenotypic features (Hasan et al., 2021).

Broadly speaking, the existing literature for the plant leaf segmentation can be classified into two categories as shown in Figure 1: conventional and deep learning-based methods. For the conventional methods, a statistical method with graph-based models (Kumar and Domnic, 2019) was proposed to segment the plant image and leaf counting, where the image enhancement techniques and the transformation from RGB to HSV were used to improve the quality of the input image. To avoid the problem of leaf over-segmentation, green channel information (Wang et al., 2018) was used to remove the background information, and the Sobel operator was improved to segment cucumber leaves. To detect the occluded plant leaves, leaf shape (Xia et al., 2013) was fused into the energy function to segment the leaf images. To deal with the complex background and the strong illumination, Larese et al. (2014) proposed a leaf vein analysis method for leguminous leaf segmentation and classification. The automatic segmentation method for plant leaf images under complex background was proposed to obtain the segmentation results. Scharr et al. (2016) uses the supervised classification with a neural network along with color and watershed transform for plant leaf segmentation and counting. Kuznichov et al. (2019) proposed a schema to augment the training dataset and remain the geometrical structure of the plant leaf by constructing a generation synthetic data. To segment multiple leaves at the same time and deal with the leaf over-segmentation, a deep extraction method for plant leaf (Amean et al., 2021) was proposed by incorporating multiple features, such as color, shape, and depth information. Lin et al. (Lin et al., 2023) proposed a self-supervised blade segmentation framework consisting of a self-supervised semantic segmentation model, a color-based blade segmentation algorithm, and a self-supervised color correction model. A self-supervised semantic segmentation model (Lin et al., 2023) was proposed to deal with the complex lighting conditions. The model was comprised of the features extracted from the CNN-based network and the fully connected Conditional Random Fields (CRFs), thus significantly reducing the impact of complex backgrounds and variations within the leaf and non-leaf regions.

In recent years, the deep learning-based method has outperformed the conventional segmentation methods and shows great potential in processing plant phenotypic tasks (Bhagat et al., 2021; Chandra et al., 2020). The SegNet-based model (Aich and Stavness, 2017) with the encoder-decoder architecture was used to segment plant leaves and leaf counting. Three RGB images and the segmentation mask of leaf counting were used as four input channels to build a regression model. Thus, the SegNet-based model can solve the problem of leaf counting (Ubbens and Stavness, 2017). To segment multiple objects, the instance segmentation model (Romera-Paredes and Torr, 2016) was proposed based on an end-to-end recurrent neural network (RNN). The model designed a spatial attention module to extract small patches, and then uses a convolutional long short-term memory (LSTM) network to build the relation of these patches. By doing so, the model can finish plant leaf segmentation and leaf counting. To solve the target occlusion problem, Ren et al. (Ren and Zemel, 2017) used an RNN-based architecture to generate continuous regions of interest and designed a human-like counting process based on the attention mechanism, thus making it a more accurate segmentation for each object in turn. Lin et al. (Lin et al., 2019) proposed a self-supervised CNN-based framework for leaf segmentation. The model first used self-contained information to classify each pixel, and then the segmentation algorithm for the color leaf images was used to identify the leaf region. Finally, a self-supervised color-based correction model was proposed to segment the complex images taken under complex lighting conditions. As shown in Table 1, we summarize the work related to plant leaf segmentation.

It is well-known that U-Net (Ronneberger et al., 2015) is one of the most efficient models and widely used for specific object extraction in image segmentation. U-Net and its variants (Shen et al., 2017) have achieved competitive performance in many computer vision tasks, such as ResU-Net (Zhang et al., 2018), U-Net++ (Zhou et al., 2019), DenseNet (Huang et al., 2017), 3D U-Net (Li et al., 2020), V-Net (Milletari et al., 2016). Bhagat et al. (2022) proposed a modified U-Net architecture for plant leaf segmentation, where an EffcientNet-B4 module was used as an encoder to extract the image feature. Meanwhile, a redesigned skip connection and the residual modules of the decoder were used to reduce computational cost. However, these methods usually ignored the global context information. To be exact, these models could not extract the long-range correlation between pixels, especially for the pixels surrounding the boundary of the objects. The effective method for obtaining the precise location and boundary of the segmentation object was to extract the global context information of the feature map and the long-range correlation between pixels. Transformer has been proved to be an efficient self-attention mechanism to establish long-term dependencies in the field of natural language processing (NLP). More recently, it was introduced into the visual classification tasks. Ramachandran et al. (Ramachandran et al., 2019) explored a novel ResNet-based model by replacing all spatial convolutional layers with the self-attention layers. However, the local self-attention might still lose part of the global structural information. In order to obtain global information of visual images, Vision Transformer (ViT) (Dosovitskiy et al., 2020) inspired by Transformer was proposed to solve the natural image recognition task. ViT first divided the image into several non-overlapping patches, and then used Transformer with the self-attention mechanism to calculate the global information between each token to obtain the global context information. To further reduce the sequence length and computational complexity, Swin

TABLE 1 The related works in plant leaf image.

Categories	Author	Method
Conventional method	Kumar and Domnic, 2019	A statistical method with graph-based models
	Wang et al., 2018	The Sobel-based model with green channel information
	Xia et al., 2013	The modified active shape models for plant leaf detection
	Larese et al., 2014	Automatic classification modle for legumes image
	Kuznichov et al., 2019	Augment dataset and the geometrical structure
	Amean et al., 2021	Self-supervised blade segmentation framework
	Lin et al., 2023	Self-supervised semantic segmentation model for complex lighting conditions
Deep learning-based method	Aich and Stavness, 2017	The SegNet-based model for leaves and leaf counting
	Ubbens and Stavness, 2017	A deep learning platform for complex plant phenotyping
	Romera-Paredes and Torr, 2016	Recurrent instance segmentation
	Ren and Zemel, 2017	End-to-end instance segmentation with recurrent attention
	Lin et al., 2019	A self-supervised CNN-based framework for leaf segmentation

Transformer (Liu et al., 2021) used a shifted window to calculate the local self-attention. By establishing a shifted window, two adjacent windows could interact with each other, and cross connections were established between the widows of the upper and lower layers, which improved the effect of global context.

To address these problems, we built a pepper leaf dataset focused on the disease detection segmentation, and propose a bidirectional attention fusion network, named BAF-Net, to obtain the for pepper leaf segmentation. BAF-Net is comprised of three parts: multi-scale fusion feature (MSFF) branch, full-scale feature fusion (FSFF) branch, and bidirectional attention feature fusion (BAF) modules. The backbone of the MSFF branch is a U-shaped network architecture. By incorporating the Swin-Transformer block and the CNN-based module, a cascaded hybrid module (Swin-Trans-Conv) is constructed, to obtain multi-scale fusion features. In the FSFF branch, we first fuse the features of the five-layer encoder from the MSFF branch. Then, the generated features pass through several convolution blocks to obtain the full-resolution feature. The BAF module adaptively fuses the output features of the MSFF and FSFF branches, generating two corresponding features for each branch. In short, the main contributions of our work are as follows:

(1) By incorporating the Swin Transformer and CNN-based modules, we build a cascaded Swin-Trans-Conv block to replace each convolutional layer of U-Net. The Swin Transformer-based module can extract the long-range dependencies while the CNN-based module is used to obtain the local image information.

(2) An FSFF branch is designed to extract detailed information and the boundaries. By incorporating the multi-scale features which are from the outputs of the encoder in the MSFF branch, the boundary information is retained. Meanwhile, the multi-layer full-scale convolution block can extract detailed information.

(3) We propose a BAF module to adaptive share the multi-scale and full-scale features, which can adaptively compute the features of two corresponding branches according to the output features of the MSFF and FSFF branches.

(4) By verifying on four dataset of pepper leaf images, the results show that our model is superior to the state-of-the-art models in terms of the evaluation indices such as IoU and F1score.

The rest of this paper is arranged as follows. Section 2 first reviews the materials including the dataset and its labeling process. Then, the proposed model including the overall architecture, the formulation of the MSFF and FSFF branches, the BAF module, and the loss function are discussed. Finally we introduce the evaluation indices. Section 3 demonstrates the experimental results and discussion. The conclusions are summarized in Section 4.

2 Materials and methods

2.1 Dataset

In our experiments, the images of pepper leaves were taken from the farm of Nanchang Academy of Agricultural Sciences in Jiangxi Province, China. We took photos for multi-view in the real natural environment from the morning to the afternoon on August 12 and 13, 2021.Pepper leaves were seriously affected by a variety of diseases during growth. Two common diseases of pepper leaf destroyed the normal growth of pepper, such as the brown spot disease and the early blight disease. Meanwhile, we also collect healthy pepper leaves to expand our dataset. As shown in Table 2, there are 3921 pepper leaf images in our dataset including the healthy pepper leaves (HPL) and two different categories of infection (2606 images): spot disease (SD) and early blight disease

TABLE 2 Four datasets for the validating the proposed model on the pepper leaf.

Dataset	Test	Training	Validation	Total
Spot Disease (SD)	186	1015	184	1385
Early Blight Disease (EBD)	164	895	162	1221
Healthy Pepper Leaf (HPL)	176	965	174	1315
Total Pepper Leaf (TPL)	526	2875	520	3921

FIGURE 1
The sample dataset with different background.

(EBD), and several examples are shown in Figure 1. As shown in Table 2, the SD, EBD, and HPL datasets contain 1385, 1221, and 1315 images. The total pepper leaf (TPL) dataset is comprised of the SD, EBD, and HPL datasets. In our experiment, the images of each image dataset are split into the training set, the validation set and the test set, and the image numbers of the training set. Meanwhile, in order to evaluate the robustness of the BAF-Net, the images were taken with different complex background as shown in Figure 1.

2.2 Dataset labeling

In the following section, we present a data labeling process, and the labeled images are used for validating the proposed model. To accurately annotate the given images, we use the open-source tool named as LabelMe[1], which was developed by the computer science and artificial intelligence laboratory of MIT university. It allows users to annotate images manually to build image dataset for image segmentation. The pixel-by-pixel way carefully delineated the boundary of each leaf. All these images in the experiment are marked using this tool. Thereafter, each annotated image generates a binary segmentation mask, where the intensity values of the foreground and background are 1 and 0, respectively. During annotating the dataset, we retain the same size as the input image. In view of the computational cost in deep learning, we set the size of the input image to 512×512.

2.3 Method

2.3.1 Overall architecture

In the field of image segmentation, U-Net has become one of the most successful network frameworks. It consists of a contracting path and an expanding path, where the contracting path is used to capture the image feature while the expanding path can achieve object localization. In each encoder-decoder layer, a skip connection layer transforms the low-level and high-level information. The model uses a convolution layer with fixed kernel size to extract image features, However, it is difficult to capture long-range semantic information. Although Transformers (Dong et al., 2019) can effectively encode the long-range dependencies, it is difficult to obtain local details and accurate boundaries of pepper leaves. To solve this problem, we propose a bidirectional attention fusion network by combining CNN and Transformer for pepper leaf segmentation, also named as BAF-Net,

1 https://github.com/wkentaro/labelme

where CNN is used to extract the local image information while the Transformer-based module can capture the long-range dependencies. As shown in Figure 2, the multi-scale branch is used to extract the global features while the full-scale feature can retain the detailed boundary information. The bidirectional fusion module is designed to concatenate the multi-scale features and the full scale features.

Specifically, BAF-Net includes three parts: a multi-scale feature fusion (MSFF) branch, a full-scale fusion feature (FSFF) branch, and bidirectional attention fusion (BAF) modules. In the MSFF branch, the network structure is similar to U-Net, composed of an encoding path and a decoding path. Different from the U-Net model, the encoder is replaced by a hybrid module by incorporating the convolutional layer and the Swin Transformer (Liu et al., 2021) module, and the decoder is composed of convolutional modules. In the FSFF branch, we first upsample four features: the output features of the encoder from the 2nd layer to the 4th layer, and the 5th layer of the decoder. Four output features are the same size as the first layer's output feature in the MSFF branch. Then, we fuse five generated features, and the generated feature is passed through four continuous convolutional modules. Each convolutional module is activated by the convolution layer, batch normalization, and the ReLU activation function. In the BAF module, the input features are from the output feature of the decoder in the MSFF branch and the output feature of the corresponding convolutional module in the FSFF branch. By incorporating the MSFF and FSFF branches, the improved model not only achieves the full resolution feature but also extracts the comprehensive and multi-scale features.

2.3.2 Multi-scale feature fusion branch

The transformer-based model (Dosovitskiy et al., 2020; Cao et al., 2021) has a more robust representation than the CNN-based model while building the long-range dependencies. In order to extract the global features, we explore a hybrid Swin-Trans-Conv block by combining the Swin-Transformer encoder and the convolutional layer, which is used to replace the convolutional layer of the encoder in the MSFF branch. As shown in Figure 3A, the backbone network including an encoder network and a decoder network is similar to U-Net. In the encoder network, we use a hybrid module by combining the convolutional layer and the Swin Transformer block, also called as Swin-Trans-Conv block, to replace each convolutional layer of U-Net, where an average pooling operator perform the downsampling process and the size of the feature maps are changed into half of the original. The decoder network is comprised of four convolutional layers and four upsampling operators. The upsampling operation is achieved by performing a deconvolutional operator with the stride of 2. The convolutional layer consists of a convolutional operator, batch normalization, and a ReLU activation layer. The number of channels in five layers corresponding to the 1st layer to the 5th layer is 32, 64, 128, 256 and 512, respectively.

Assuming that the input feature is $X \in R^{B\times H\times W\times C}$, where B, C, H and W represent the batch size, the channel number, and the image height and width of the input feature, respectively. In the Swin-Trans-Conv block as shown in Figure 3A, we first transforms the input feature X into $X' \in R^{B\times H\times W\times C}$. Then, we perform a 1×1 convolution operator on the generated feature, and split the generated feature Y into two groups F_{Trans} and F_{conv}, which can be expressed as:

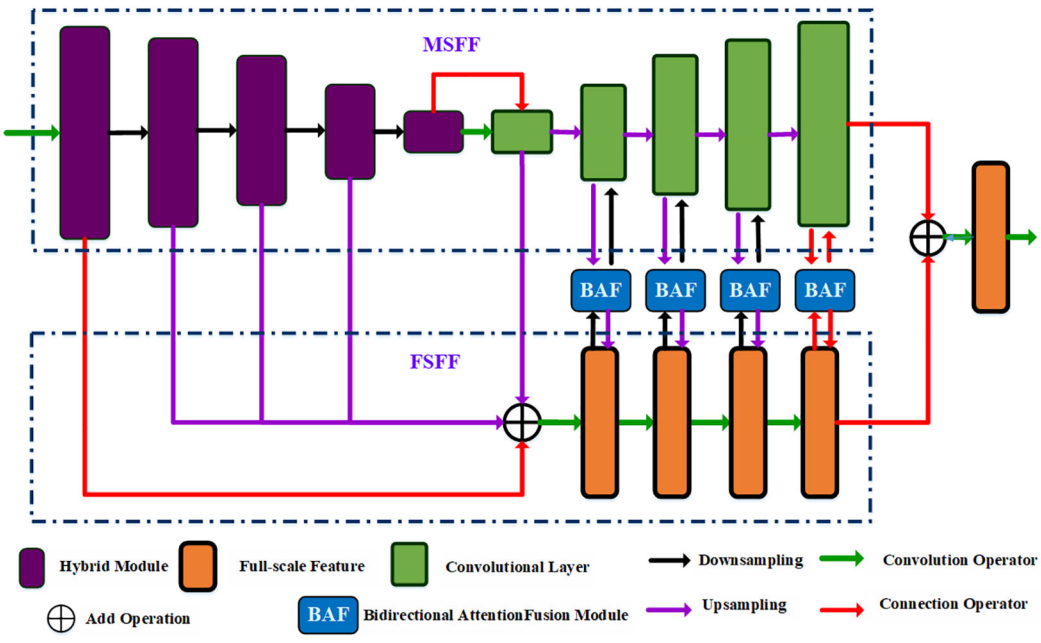

FIGURE 2
The overall framework of the proposed BAF-Net, which includes three main modules such as the multiscale feature fusion branch, GAM and decoders, where the decoder includes the global context module (GAM) and FAM with LAM.

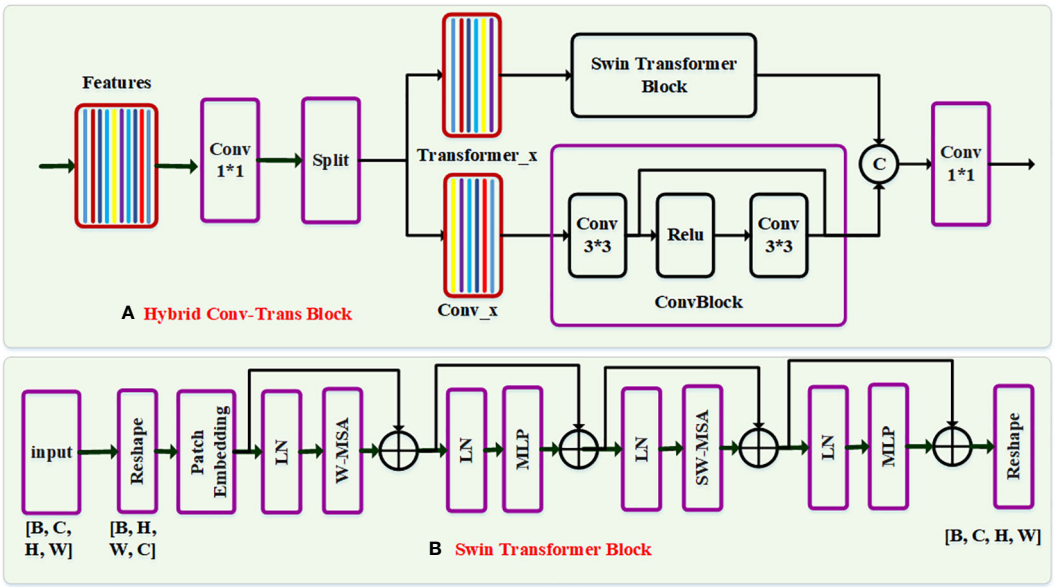

FIGURE 3
The network structure of the Swin-Trans-Conv block. In each block, the input feature is first passed through a 1×1 convolution, and subsequently is split evenly into two feature map groups, each of which is then fed into a Swin transformer block and a residual 3×3 convolutional (RConv) block, respectively. Afterwards, the output features of the Swin-Trans-Conv block and the RConv block are concatenated and then passed through a 1×1 convolution to generate a novel feature *via* a residual path.

$$Y = \mathrm{Conv}_{1\times1}(\mathrm{Re\,shape}(X)) \quad X, Y \in R^{B\times C\times H;\times W} \tag{1}$$

$$F_{trans},\ F_{conv} = \mathrm{Split}(Y) \quad F_{Trans}, F_{conv} \in R^{B\times C/2\times H\times W} \tag{2}$$

where $\mathrm{Re\,shape}(\cdot)$ is a reshape operator on two feature matrix F_{trans} and F_{conv}, $\mathrm{Conv}_{1\times1}(\cdot)$ denotes a 1×1 convolutional operator, and $\mathrm{Split}(\cdot)$ represents a split operation on the multidimensional matrix. Finally, the feature F_{Trans} is passed through a module based on Swin Transformer (Swin-Trans) encoder, and the generated feature map F'_{trans} is written as:

$$F'_{trans} = \mathrm{SwinTrans}(F_{trans}) \tag{3}$$

Similarly, the feature map F_{conv} passes through a residual convolution module, and the generated feature F'_{conv} is defined as:

$$F'_{conv} = \mathrm{RConv}(F_{conv}) \tag{4}$$

where $\mathrm{RConv}(\cdot)$ is the residual convolution module, which is comprised of a 3×3 convolution filter, a ReLU activation layer, and a 3×3 convolution filter by a residual path, which is rewritten as:

$$F^3 = \mathrm{Conv}_{3\times3}(F_{conv}) \tag{5}$$

$$F'_{conv} = \mathrm{Conv}_{3\times3}(\mathrm{Re\,lu}(F^3)) + F^3 \tag{6}$$

where F^3 is the feature map performed a 3×3 convolution operation on the feature map F_{conv}, $\mathrm{Conv}_{3\times3}$ is a 3×3 convolutional layer, and $\mathrm{Re\,lu}(\cdot)$ is a ReLU activation layer. Finally, we concatenate two features F'_{trans} and F'_{conv}, and then perform a 1×1 convolution filter by a residual path, which is represented as:

$$X_{out} = \mathrm{Conv}_{1\times1}(F'_{trans} \odot F'_{conv}) + X \tag{7}$$

where $\odot$ denotes the concatenation operation.

Meanwhile, to construct the Swin-Trans module as shown in Figure 3B, the input feature F_{Trans} is split into small patches, and each patch size is set to $P \times P \times P$, where P is a positive integer and the number of the patches is $S=[H/P]\times[W/P]\times[C/P]$. For the feature F^i_{Trans} of the i-th layer with the 3D patches, we first compute the multi-head self-attention in a small window (W-MSA), which can be formulated as:

$$\hat{F}^i_{out} = \text{W-MSA}(\mathrm{LN}(F^i_{Trans})) + F^i_{Trans} \tag{8}$$

$$\hat{F}^i_{mlp} = \mathrm{MLP}(\mathrm{LN}(\hat{F}^i_{Out})) + \hat{F}^i_{Out} \tag{9}$$

where W-MSA$(\cdot)$ denotes the window multi-head self-attention, $\mathrm{LN}(\cdot)$ is the layer normalization operator, and $\mathrm{MLP}(\cdot)$ denotes a multilayer perceptron module with two fully-connected layers and the GELU activation function. Then, the generated feature $\hat{F}^i_{mlp}$ is passed through the multi-head self-attention in the shifted window (SW-MSA), which is represented as:

$$\hat{F}^i_{sw} = \text{SW-MSA}(\mathrm{LN}(\hat{F}^i_{mlp})) + \hat{F}^i_{mlp} \tag{10}$$

$$F^i_{out} = \mathrm{MLP}(\mathrm{LN}(\hat{F}^i_{sw})) + \hat{F}^i_{sw} \tag{11}$$

where SW-MSA$(\cdot)$ denotes the shifted window multi-head self-attention, and F^i_{out} is the output feature of the i-th layer.

Finally, the output feature F^{i}_{out} is reshaped into the same size of the input feature in the Swin-Trans module.

It is worth noting that the Swin-Trans-Conv block has several advantages. First, it integrates the local modeling capability of the convolution module and the global modeling capability of the Swin-Trans module. Secondly, the split and concatenation operations are used for two branches to extract different features, reducing the computational complexity and the number of parameters.

2.3.3 Full-scale feature fusion branch

The edge and detailed image information may be lost in the U-shape network framework due to the continuous downsampling operators. To solve this problem, we design an MSFF branch to retain the detailed information, and the network structure is shown in Figure 2B. We fuse the output features of the first 1st to 4th layer in the encoder of the MSFF branch and the output feature of the decoder of the 5th layer in the decoder since the multi-scale features can enhance the edge information (Liu et al., 2023). For four output features from the MSFF branch, we first carry out a 1×1 convolution filter to reduce the channel number. Then, we perform the upsampling operator on the four features, and the four generated features have the same size with the first channel feature. Then, we integrate four generated features into the input feature by a residual path, which can be expressed as:

$$X_i^{up} = \mathrm{Conv}_{1\times1}\left(\underbrace{\mathrm{up}\left(\cdots \mathrm{up}(X_i)\right)}_{i-1}\right) \qquad i = 2, \cdots, 5 \tag{12}$$

$$X_{fuse} = \sum_{i=2}^{5} X_i^{up} + X \tag{13}$$

where $\mathrm{Conv}_{3\times3}$ denotes a 3×3 convolutional filter.

Finally, the novel feature passes through four continuous convolutional modules. Each convolutional module includes a 3×3 convolution layer, batch normalization, and a ReLU activation layer. To reduce the computational cost and the parameters, we keep each channel number of four features equal to that of the first layer in the MSFF branch. In this paper, the channel number is set at 32. The operations for each convolution block are presented as follows:

$$\begin{cases} X_f^4 = \mathrm{Re\,lu}(\mathrm{BN}(\mathrm{Conv}_{3\times3}(f_{fuse}))) & i = 5 \\ X_f^{i-1} = \mathrm{Re\,lu}(\mathrm{BN}(\mathrm{Conv}_{3\times3}(X_f^i))) & i = 2, \cdots, 4 \end{cases} \tag{14}$$

2.3.4 Bidirectional attention fusion module

In order to achieve the multi-scale and full-scale features, we designed a BAF module to generate the corresponding output features for the MSFF and FSFF branches. As shown in Figure 4, the BAF module includes multi-scale feature guidance (MSFG) module and full-scale feature guidance (FSGM) module. For the MSGM module, we first conduct the downsampling operation on the input feature of the FSGM module, and the novel feature maps have the same spatial dimensions with the same with that of the MSGF map, which can be expressed as:

$$F_{fg}^{dn} = \mathrm{DN}(F_{fg}) \tag{15}$$

where $\mathrm{DN}(\cdot)$ denotes the downsampling operation. Then, we concatenate the output feature of the MSFF branch F_{ms} and the feature F_{fg}^{dn}, and perform a 1×1 convolution module on the novel feature map to compress the number of channels, we can obtain the feature map:

$$F_{ms}^{c} = \mathrm{Conv}_{1\times1}(F_{ms}\ \mathrm{e}\ F_{fg}^{dn}) \tag{16}$$

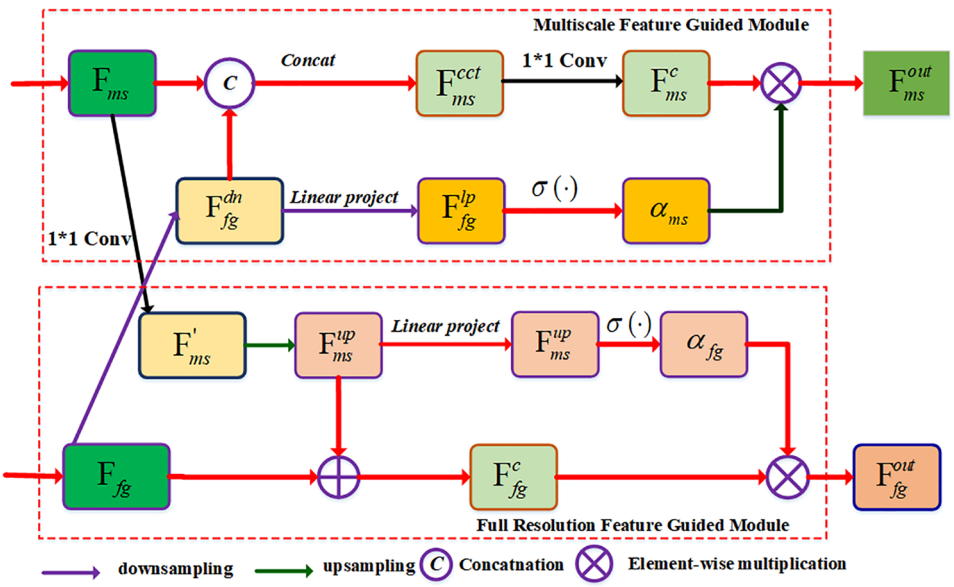

FIGURE 4
The network structure of the BAF module. Two input features F_{ms} and F_{fg} are from the output features of the MSFF and FSFF branches, respectively. The BAF module contains a multi-scale feature guided (MSFG) module and a full resolution feature guided (FRFG) module. The MSFG module is used to generate the multi-scale feature while the FRFG module is used to generate the full-scale feature.

At the same time, we project the feature F_{ms}^{c} to compress the feature map into a channel along the channel direction, and use the Sigmoid activation function to obtain the global attention map, which is defined as:

$$\alpha_{ms} = \sigma_{sig}(\Pr oj(F_{ms}^{c})) \tag{17}$$

where $\Pr oj(\cdot)$ denotes the linear projection function, $\sigma_{sig}(\cdot)$ denotes the Sigmoid function, and $\alpha_{ms} \in [0,1]$ is the spatial attention map of the feature F_{ms}^{c}. It is obvious that the spatial attention map α_{ms} calculates the spatial weight of each pixel, and the calibrated feature map is expressed as:

$$F_{ms}^{out} = \alpha_{ms} \otimes F_{ms}^{c} \tag{18}$$

Finally, the feature F_{ms}^{out} is transformed to the next convolution layer.

In the FSFG module, we first perform a 1×1 convolutional filter on the multi-scale feature map F_{ms} to compress the number of channels. The expression is as follows:

$$F_{ms}^{'} = Conv_{1\times1}(F_{ms}) \tag{19}$$

Then, we upsample the multi-scale feature map to make the generated features have the same spatial dimension as that of the full-scale feature. The expression is as follows:

$$F_{ms}^{up} = up(F_{ms}^{'}) \tag{20}$$

where $up(\cdot)$ denotes the upsampling operator. Afterwards, two features F_{ms}^{up} and F_{fg} are fed into the convolutional layer to generate a new feature F_{fg}^{c}, which is written as:

$$F_{fg}^{c} = Conv_{3\times3}(F_{ms}^{up} \oplus F_{fg}) \tag{21}$$

where $\oplus$ represents the pixel-wise addition operation. Meanwhile, we use linear projection to compress the feature into a channel along the channel direction, and then use the sigmoid activation function to obtain the global attention map:

$$\alpha_{fg} = \sigma_{sig}(\Pr oj(F_{fg}^{c})) \tag{22}$$

where $\alpha_{fg} \in [0, 1]$ is the spatial attention map of F_{fg}^{c}, which is used to calculate the spatial position weight of each pixel. The calibrated feature map can be represented as:

$$F_{fg}^{out} = \alpha_{fg} \otimes F_{fg}^{c} \tag{23}$$

Finally, it is input to the convolution layer of the next FSFG module.

2.3.5 Training loss

The network should be trained to obtain the best training parameters. It is known that the loss function is essential to the predicted performance of the segmentation model. The loss function is used to measure the deviation between the model prediction and the ground truth. The binary cross entropy (BCE) is a loss function widely used in binary image segmentation tasks. Assuming that the input predicted result is p, and the corresponding ground truth label is g, the BCE loss function is defined as:

$$L_{bce}(p, \quad g) = -\sum_{i=1}^{N}[g_x \log(p_x) + (1 - g_x)\log(1 - p_x)] \tag{24}$$

The intersection over union (IoU) loss is defined as:

$$L_{IoU}(p, \quad g) = -\log\left(\frac{\sum_{i=1}^{N}|g_x \cdot p_x|}{\sum_{i=1}^{N}(g_x + p_x - |g_x \cdot p_x|)}\right) \tag{25}$$

Therefore, our final loss includes L_{bce} and L_{IoU}, which can be expressed as:

$$L_{total}(p, \quad g) = \alpha L_{bec}(p, g) + (1 - \alpha)L_{IoU}(p, g) \tag{26}$$

The weight α is a coefficient to balance the importance of two loss functions, and we set α=0.5.

2.4 Performance evaluation

In order to verify the segmentation performance, we use six evaluation indices to evaluate the accuracy of the model on the pepper leaf datasets. Six evaluation indices include: pixel accuracy (PA), pixel recall (PR), pixel precision (PP), pixel specificity (PS), intersection over union (IoU) and F1 score. We assume that TP (True Positive) represents the number of pixels that are both 1 in the predicted value and the label value, TN (True Negative) represents the number of pixels that are both 0 in the predicted value and the label value, FP (False Positive) represents the number of pixels that are 1 in the predicted value and 0 in the label value, and FN (False Negative) represents the number of pixels that are 0 in the predicted value and 1 in the label value. The expression of the pixel accuracy is written as follows:

$$PA = \frac{TP + TN}{TP + TN + FP + FN} \tag{27}$$

PR is defined as follows:

$$PR = \frac{TP}{TP + FN} \tag{28}$$

PP is defined as follows:

$$PP = \frac{TP}{TP + FP} \tag{29}$$

F1 score is defined as:

$$F1 = \frac{2 \times PR \cdot PP}{PR + PP} \tag{30}$$

PS is defined as follows:

$$PS = \frac{TN}{TN + FP} \tag{31}$$

From Equations (27)-(31) and the IoU as defined in Equation (25), it can be seen that six evaluation indices range from 0 to 1. The higher the index values are, the best segmentation performance is obtained. Generally speaking, the mean IoU (mIoU) is used to evaluate the segmentation performance on a given dataset.

3 Experiments

In this section, we present the experimental results including the experimental settings, the comparison with the state-of-the-arts models, the ablation study and the discussion.

3.1 Experimental settings

All models in the experiment are carried out on Intel (R) Core (R) i7-8700K CPU 3.70GHz CPU and Nvidia GeForce TITAN XP 12 GB GPU with 48G RAM. The programs are conducted on the Ubuntu 16.04 with the Conda environment. In the BAF-Net, the parameter settings are as follows: the batch size is set to 4, the number of iterations (epoch) is set to 60, and each epoch contains 350 batches. During the training process, the network is optimized using stochastic gradient descent (SGD), the initial learning rate is set to 0.01.

3.2 Comparison with the state-of-the-arts models

We compared BAF-Net with the state-of-the-art methods on four pepper leaf datasets, such as the SD, EBD, HPL, and TPL datasets. For fairness, these models are running on the same training dataset, the validation dataset, and the test dataset. The comparative models on the pepper leaf dataset involve U-Net (Ronneberger et al., 2015), AttU-Net (Oktay et al., 2018), Swin-UNet (Cao et al., 2021), SCUNet (Zhang et al., 2022) and the proposed BAF-Net. We set the training epochs to 60 for each trained model.

Table 3 shows the test results on the SD dataset using five different state-of-the-art models. Compared with U-Net, the proposed model has a precision increase of 7.48%, IoU increase of 3.88%, and F1 score increase of 5.0%. It also shows that PA score has the relative improvement of 0.5% on the SD dataset. For the attention U-Net model, the segmentation results on five indices are close to that of the U-Net. In addition, Swin-UNet and SCUNet have the similar segmentation performance. However, the segmentation performance of U-Net exceeds two models in terms of six evaluation indices. The reason is that Swin-UNet and SCUNet containing the transformer-based modules attain better segmentation results only if more efficient pre-trained model is provided. From Table 3, where the highest score for each indicator is shown in bold, our model can obtain the best segmentation performance in terms of five evaluation indices including PA, PP, PS, IoU, and F1 scores compared with other models. By evaluating the segmentation performance of five models, we also give several examples of the segmentation results using these compared methods as shown in Figure 5.

Table 4 presents the segmentation results of five segmentation models on EBD pepper leaf dataset, in which the highest score of each index is shown in bold. From the experimental results, the proposed model has the highest scores among the six indices including PA, PR, PP, PS, IoU and F1 scores. Specifically, compared with U-Net, BAF-Net increased PA by 0.62%, PR by 0.06%, PP by 3.44%, PS by 1.7%, IoU by 5%, and F1 score by 7.04%. Attention U-Net is only lower than BAF-Net in terms of the indices IoU and F1-score, with a decrease of 4.92% and 6.59%, respectively. Compared with Swin-Unet and SCUNet, the proposed model has significant improvement in terms of six indices. The proposed BAF-Net have significant improvement in terms of PP, reaching the increase by 7.25% and 14.29%, respectively. By evaluating the segmentation performance of five deep learning-based models, we find these models can obtain better segmentation results than the traditional methods. Meanwhile, we also give the examples of the segmentation results using these compared methods as shown in Figure 6.

Table 5 shows the validation results of five different models on the HPL data set, with the highest score for each indicator shown in bold. From the experimental results, the proposed model can obtain the best segmentation accuracy in terms of PA, PP, PS, mIoU and F1 score. Compared with U-Net, the proposed model has increased PA by 0.21%, PP by 1.51%, PS by 1.52%, IoU by 0.01%, and F1 score by 0.27%. The attention U-Net has the similar segmentation results with U-Net. Our model has significant improvement than Swin-UNet and SCUNet in terms of the PP, mIoU and F1 score. Compared with the Swin-UNet, the PP, mIoU and F1 scores have increased by 4.95%, 3.60% and 2.43%, respectively. Compared with the SCUNet, the PP, mIoU and F1 score has increased by 2.51%, 2.22% and 1.58%, respectively. Meanwhile, we also give the example of the segmentation results for qualitative comparison, and the representative examples are shown in Figure 7.

The experimental results on the TPL dataset are shown in Table 6, with the highest score in each indicator represented in bold. It can be seen that our model obtains the best segmentation results in terms of the six indices among five models. Compared with U-Net, the proposed model has IoU increased by 0.01%, PA increased by 0.13%, PR increased by 0.03%, and PP increased by 0.87%. The PS score is 0.01% higher than that of U-Net, and F1 score is 0.48%

TABLE 3 The segmentation results on the SD dataset using five different models.

Model	PA(%)	PR(%)	PP(%)	PS(%)	mIoU(%)	F1(%)
UNet	98.70	**99.09**	91.10	98.65	91.80	94.93
Attention U-Net	98.24	98.11	88.72	98.26	90.09	93.18
Swin-UNet	97.68	97.90	85.36	97.65	86.76	91.20
SCUNet	97.42	98.14	83.68	97.32	87.34	90.33
Ours	**99.20**	98.74	**98.58**	**99.80**	**95.68**	**96.75**

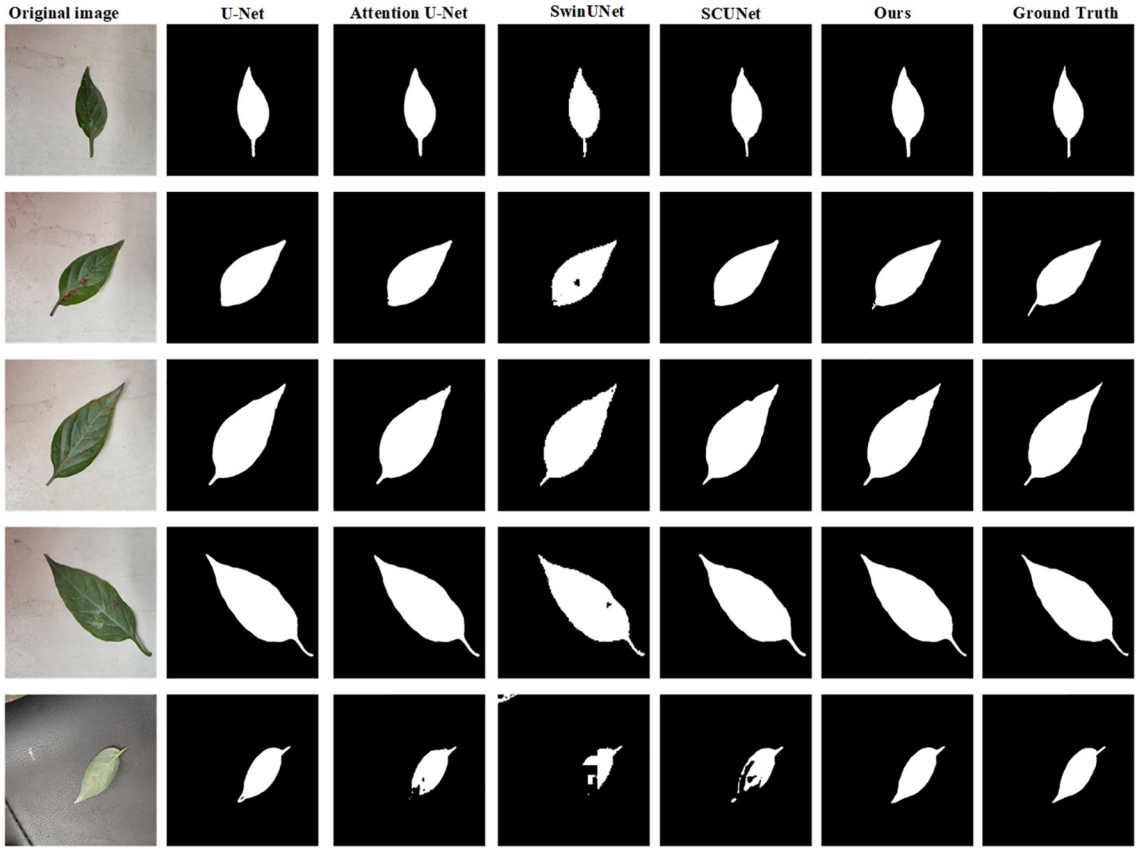

FIGURE 5
Examples of the predicted results using five different models on the SD dataset. From the 1st column to 7th column: the original images, the predicted results using the U-Net, attention U-Net, Swin-UNet, SCUNet, the proposed model and ground truth (GT), respectively.

higher than that of U-Net. Compared with attention U-Net, the proposed model has significant improvement in terms of six indices. However, Swin-UNet and SCUNet do not improve the segmentation results compared with the U-Net. In summary, BAF-Net has obvious advantages in segmenting the pepper leaf from the natural images.

3.3 Ablation study

In this section, we perform an ablation study to validate the effectiveness of each module. Especially, we consider the basic U-Net architecture as the baseline, namely the simple U-Net (SU-Net), which is similar to U-Net with half of the channel number of U-Net. In the ablation experiments, we take SU-Net, MFF, MRF, and BAF as four basic modules. Our experimental strategy is to add a module each time, and it is proven to be effective. We approve that it is effective in subsequent studies. Strictly speaking, we selected four unique models, such as SU-Net, SU-Net-MFF, SU-Net- MFF-MRF, and BAF-Net, to verify that different modules are still valid when each model is added to SU-Net each time.

As shown in Table 7, we first experiment SU-Net-MSFF by replacing the convolution layer of the encoder in the SU-Net model with the Swin-Trans-Conv block, which is formulated by adding the MSFF module into SU-Net. Experiments show that PA, PR, PP, PS, mIoU and F1 score of the SU-Net-MSFF model are 98.94%, 96.87%, 96.90%, 99.36%, 95.63% and 96.88, respectively. Then, by adding the FSFF module to SU-Net-MSFF, the results show that the PA,

TABLE 4 The segmentation results on the EBD dataset using different models.

Model	PA(%)	PR(%)	PP(%)	PS(%)	mIoU(%)	F1 (%)
UNet	96.98	98.11	81.54	95.82	81.76	84.06
Attention U-Net	95.57	96.63	75.08	95.41	81.84	84.51
Swin-Unet	96.02	95.55	77.73	96.08	80.63	85.72
SCUNet	94.49	95.70	70.69	94.32	77.33	81.31
BAF-Net(ours)	**97.60**	**98.17**	**84.98**	**97.52**	**86.76**	**91.10**

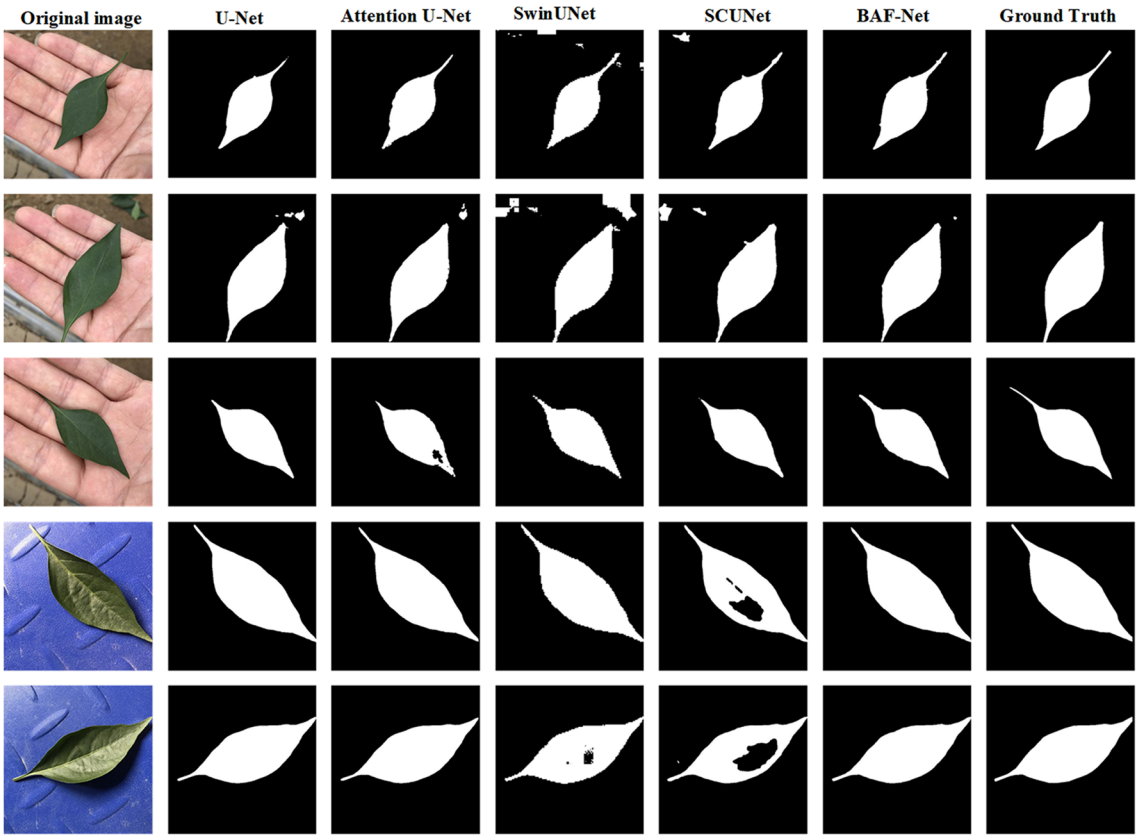

FIGURE 6
Examples of the predicted results using five different model on the EBD dataset. From the 1st column to 7th column: the original image, the predicted results using the U-Net, attention U-Net, Swin-UNet, SCUNet, and our model, respectively.

PA, PP, PS, IoU and F1 scores of the SU-Net-MSFF-FSFF are 98.94%, 96.62%, 97.14%, 99.42%, 95.68% and 96.98%, respectively. Compared with the SU-Net-MSFF, PR, PS, IoU and F1 score of the SU-Net-MSFF-FSFF model are increased by 0.24%, 0.06%, 0.05% and 0.10%, respectively. Finally, we experiment BAF-Net by fusing the output features of the decoder in the MSFF and FSFF branches to the BAF modules. The results show that PA, PR, PP, PS, IoU and F1 score of BAF-Net are 98.98%, 6.82%, 97.20%, 99.43%, 95.86% and 97.01%, respectively, which are increased by 0.04%, 0.2%, 0.06%, 0.01%, 0.18% and 0.03%, respectively. From the segmentation results, we can see that the addition of the Swin-Trans-Conv block expands the receptive field and enhances the feature extraction ability of SU-Net, enabling it to obtain different levels of information at the same time. The full-resolution features enable the proposed model to retain image local details. By combining multi-scale information and full scale information, it can extract deeper structural information. Therefore, the combination of the three modules can obtain the best performance.

3.4 Discussion

The above analysis shows that the segmentation results of these deep learning-based segmentation models are suitable. Compared with the classical methods based on the variational statistics theory (Costa et al., 2019; Fang et al., 2019a; Fang et al., 2019b; Gao and Lin, 2019; Liu et al., 2020; Fang et al., 2021a; Fang et al., 2021b; Liu et al., 2021; Ward et al., 2021), the deep-learning-based models can obviously obtain

TABLE 5 The segmentation results on the HPL dataset using different models.

Model	PA(%)	PR(%)	PP(%)	PS(%)	mIoU(%)	F1 (%)
U-Net	99.09	**98.51**	95.84	99.02	96.11	97.07
Attention U-Net	99.21	97.69	96.34	99.44	95.53	97.01
Swin-UNet	98.63	97.57	92.40	98.79	92.52	94.91
SCUNet	98.88	96.69	94.84	99.21	93.90	95.76
BAF-Net(ours)	**99.30**	97.32	**97.35**	**99.60**	**96.12**	**97.34**

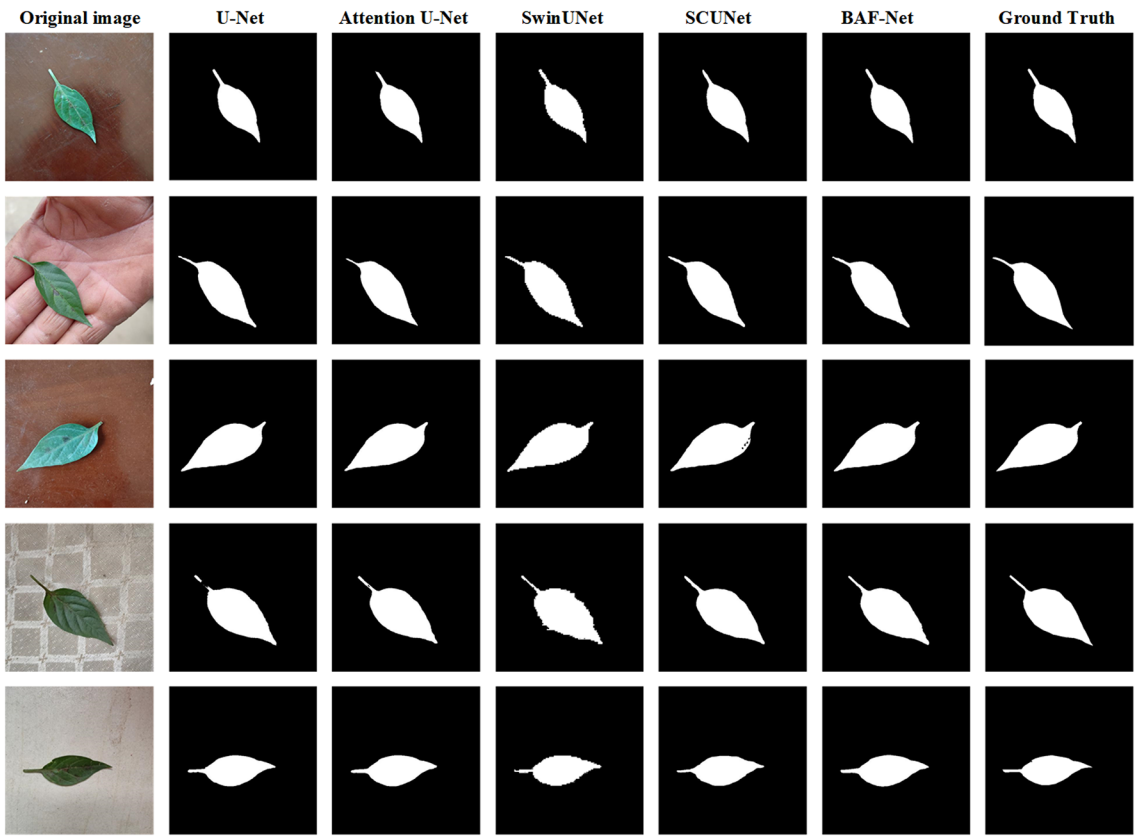

FIGURE 7
Examples of the predicted results using five different model on the HPL dataset. From the 1st column to 7th column: the original image, the predicted results using the U-Net, attention U-Net, Swin-UNet, SCUNet, and our model, respectively.

better classification results. In our work, to capture the long-range dependencies between different pixels, we propose a bidirectional adaptive attention fusion network called BAF-Net by exploring an adaptive attention mechanism to extract multi-scale and full-scale features simultaneously. Specifically, we first design an MSFF branch based on the encoder-decoder structure, which can not only extract local information of the target, but also learn the spatial attention to increase the receptive field. To further retain the boundary information of the segmented object, we propose a FSFF branch, and design adaptive bidirectional attention modules to achieve the bidirectional connection between the MSFF module and the FSFF module.

The results of the ablation experiment in Table 7 shows that progressive network such as SU-Net, SU-Net-MSFF, SU-Net-MSFF-FSFF and BAF-Net can improve the predicted performance of the baseline (SU-Net). Compared with the baseline, three models by progressively adding the MSFF, FSFF and BAF modules increase mIoU by 0.28%, 0.33% and 0.51%, respectively, and F1 score increased by 0.02%, 0.02% and 0.36%, respectively. From the segmentation results, it can be seen that BAF-Net has achieved the best performance. Compared with the baseline, the mIoU and F1 score of BAF-Net reaches 95.86% and 97.01%, respectively.

Although the proposed BAF-Net can obtain better performance on the four pepper leaf datasets, there are disadvantages in this work. (1) In the training process, the epoch number in our model is set to 60. Therefore, we need explore a schema to stop the training process for the deep learning-based model automatically. (2) Our

TABLE 6 The segmentation results on the pepper leaf dataset using different models.

Model	PA(%)	PR(%)	PP(%)	PS(%)	mIoU(%)	F1 (%)
U-Net	98.40	98.59	89.70	98.37	91.43	93.94
Attention U-Net	97.73	97.51	86.28	97.76	89.31	91.55
Swin-UNet	97.48	97.06	85.07	97.54	86.87	90.67
SCUNet	97.00	96.88	82.40	97.01	86.32	89.06
BAF-Net(ours)	**98.53**	**98.62**	**90.57**	**98.52**	**91.44**	**94.42**

TABLE 7 Comparison of pepper segmentation results of four models on the dataset.

Model	MSFF	FSFF	BAF	PA%	PR%	PP%	PS%	IoU%	F1%
SU-Net				98.86	96.52	96.80	99.35	95.35	96.66
SU-Net-MSFF	√			98.94	**96.87**	96.90	99.36	95.63	96.88
SU-Net_MSFF-FSFF	√	√		98.94	96.62	97.14	99.42	95.68	96.98
BAF-Net	√	√	√	**98.98**	96.82	**97.20**	**99.43**	**95.86**	**97.01**

model is supervised learning, which requires many training samples. Accordingly, in our future work, we will focus on the semi-supervised or self-supervised segmentation methods to reduce the requirements for training samples.

4 Conclusion

In our work, we propose a bidirectional adaptive attention fusion network for automatic segmentation of pepper leaves. The proposed model consists of the MSFF branch with the like-U-Net network structure, the FSFF branch, and the BAF modules with an adaptive attention mechanism. This MSFF branch fuses the Swin-Transformer-based and CNN-based modules to construct the Swin-Trans-Conv block, which replaces the convolution layer of the encoder of U-Net to expand the receptive field. In the MSFF branch, the CNN-based layer can extract the local image features while the Swin-Transformer-based module is used to extract the long-range dependencies of the channel and spatial information to expand receptive field. The FSFF branch performs multiple convolution layers keeping the same size with the original image, which is used to retain the boundary information and detail information of the segmented object. In addition, the BAF modules are used to fuse the output features of the MSFF and FSFF branch, which output the corresponding features for each branch. Compared with the existing model, our model obtain the highest evaluation indices on four pepper leaf datasets. In addition, the ablation experiment shows that the proposed three modules including MSFF, FSFF and BAF are effective. In the future, we will explore a weak-supervised model for pepper leaf segmentation since the small dataset may cause over-segmentation. Meanwhile, we study the construction of loss function and the method for augmentation dataset.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author.

Author contributions

JF: conceptualization, methodology, experiment, and writing. HL: supervision and writing- review & editing. HJ: experiment. YF: methodology and investigation. SZ: methodology and investigation. LS: experiment. XH: writing-review & editing. JL: review & editing. MG: review & editing. All authors contributed to the article and approved the submitted version.

Funding

The research described in this paper was funded by the National Natural Science Foundation of China (No. 61966001, No.62206195, No. 61866001, No. 62163004, No. 61963002, and No. 62206195), the Joint Funds of the Zhejiang Provincial Natural Science Foundation of China (No. LZY23F050001), Natural Science Foundation of Jiangxi Province (No. 20202BABL214032 and No. 20202BABL203035), Science and Technology Plan Project of Taizhou City (No. 22ywa58 and No. 22nya18), Jiangxi Engineering Laboratory on Radioactive Geoscience and Big Data Technology (No. JELRGBDT202201), the Engineering Research Center of Development and Management for Low to Ultra-Low Permeability Oil & Gas Reservoirs in West China(No. KFJJ-XB-2020-1), and the Open Fund of Key Laboratory of Exploration Technologies for Oil and Gas Resources (No. K2021-02).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Aich, S., and Stavness, I. (2017). "Leaf counting with deep convolutional and deconvolutional networks," in *Proceedings of the IEEE international conference on computer vision workshops*, 2080–2089. doi: 10.1109/ICCVW.2017.244

Amean, Z. M., Low, T., and Hancock, N. (2021). Automatic leaf segmentation and overlapping leaf separation using stereo vision. *Array* 12, 100099. doi: 10.1016/j.array.2021.100099

Bhagat, S., Kokare, M., Haswani, V., Hambarde, P., and Kamble, R. (2021). "Wheatnet-lite: a novel light weight network for wheat head detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 1332–1341. doi: 10.1109/ICCVW54120.2021.00154

Bhagat, S., Kokare, M., Haswani, V., Hambarde., P., and Kamble, R. (2022). Eff-UNet ++: A novel architecture for plant leaf segmentation and counting. *Ecol. Inf.* 68, 101583. doi: 10.1016/j.ecoinf.2022.101583

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., et al. (2021). Swin-unet: Unet-like pure transformer for medical image segmentation. In: *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III* Cham: Springer Nature Switzerland, 205–218. doi: 10.48550/arXiv.2105.05537

Chandra, A. L., Desai, S. V., Guo, W., and Balasubramanian, V. N. (2020). Computer vision with deep learning for plant phenotyping in agriculture: a survey. *arXiv*. doi: 10.48550/arXiv.2006.11391

Costa, C., Schurr, U., Loreto, F., Menesatti, P., and Carpentier, S. (2019). Plant phenotyping research trends, a science mapping approach. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01933

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., et al. (2019). "Unified language model pre-training for natural language understanding and generation", *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* 13063–13075. doi: 10.48550/arXiv.1905.03197

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner,, et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*. doi: 10.48550/arXiv.2010.11929

Fang, J., Liu, H., Zhang, L., Liu, J., and Liu, H. (2019a). Fuzzy region-based active contours driven by weighting global and local fitting energy. *IEEE Acess* 7, 184518–184536. doi: 10.1109/ACCESS.2019.2909981

Fang, J., Liu, H., Zhang, L., Liu, J., and Liu, H. (2019b). Active contour driven by weighted hybrid signed pressure force for image segmentation. *IEEE Acess* 7, 97492–97504. doi: 10.1109/ACCESS.2019.2929659

Fang, J., Liu, H., Zhang, L., Liu, J., and Liu, H. (2021a). Region-edge-based active contours driven by hybrid and local fuzzy region-based energy for image segmentation. *Inf. Sci.* 546 (6), 397–419. doi: 10.1016/j.ins.2020.08.078

Fang, J., Liu, H., Zhang, L., Liu, J., Zhou, H., and Liu, H. (2021b). Fuzzy region-based active contour driven by global and local fitting energy for image segmentation. *Appl. Soft Comp.* 100, 106982:1–16. doi: 10.1016/j.asoc.2020.106982

Gao, L., and Lin, X. (2019). Fully automatic segmentation method for medicinal plant leaf images in complex background. *Comp. Elec. Agric.* 164, 104924. doi: 10.1016/j.compag.2019.104924

Hasan, A. S. M. M., Sohel, F., Diepeveen, D., Laga, H., and Jones, M. G. K. (2021). A survey of deep learning techniques for weed detection from images. *Comput. Electron. Agric.* 184, 106067. doi: 10.1016/j.compag.2021.106067

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 4700–4708. doi: 10.48550/arXiv.1608.06993

Koirala, A., Walsh, K. B., Wang, Z., and McCarthy, C. (2019). Deep learning-method overview and review of use for fruit detection and yield estimation. *Comp. Elec. Agric.* 162, 219–234. doi: 10.1016/j.compag.2019.04.017

Kumar, P., and Domnic, S. (2019). Image based leaf segmentation and counting in rosette plants. *Inf. Proce. Agric.* 6 (2), 233–246. doi: 10.1016/j.inpa.2018.09.005

Kuznichov, D., Zvirin, A., Honen, Y., and Kimmel, R. (2019). "Data augmentation for leaf segmentation and counting tasks in rosette plants,", *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* Long Beach, CA, USA, 2019, pp. 2580–2589. doi: 10.1109/CVPRW.2019.00314

Larese, M. G., Namías, R., Craviotto, R. M., Arango, M. R., Gallo, C., and Granitto, P. M. (2014). Automatic classification of legumes using leaf vein image features. *Patt. Recog.* 47, 158–168. doi: 10.1016/j.patcog.2013.06.012

Li, Z., Pan, J., Wu, H., Wen, Z., and Qin, J. (2020). "Memory-efficient automatic kidney and tumor segmentation based on non-local context guided 3D U-net," in *International conference on medical image computing and computer-assisted intervention (MICCAI)* (Springer), 197–206. doi: 10.1007/978-3-030-59719-1_20

Lin, K., Gong, L., Huang, Y., Liu, C., and Pan, J. (2019). Deep learning-based segmentation and quantifcation of cucumber powdery mildew using convolutional neural network. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00155

Lin, X., Li, C. T., Adams, S., Kouzani, A., Jiang, R., He, L., et al. (2023). Self-supervised leaf segmentation under complex lighting conditions. *Patt. Recog* 135, 109021. doi: 10.1016/j.patcog.2022.109021

Liu, H., Fang, J., Zhang, Z., and Lin, Y. (2020). A novel active contour model guided by global and local signed energy-based pressure force. *IEEE Access* 8, 59412–59426. doi: 10.1109/ACCESS.2020.2981596

Liu, H., Fang, J., Zhang, Z., and Lin, Y. (2021). Localised edge-region-based active contour for medical image segmentation. *IET Image Process.* 15 (7), 1567–1582. doi: 10.1049/ipr2.12126

Liu, H., Fu, Y., Zhang, S., Liu, J., and Fang, J. (2022). Active contour driven by adaptive-scale local-energy signed pressure force function based on bias correction for medical image segmentation. *IET Image Process.* 16 (14), 3929–3947. doi: 10.1049/ipr2.12604

Liu, H., Fu, Y., Zhang, S., Liu, J., Wang, Y., and Wang, G. (2023). GCHA-net: Global context and hybrid attention network for automatic liver segmentation. *Comput. Biol. Med.* 152, 106352. doi: 10.1016/j.compbiomed.2022.106352

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022). In *Proceedings of the IEEE/CVF international conference on computer vision. arXiv*, 10012–10022. doi: 10.48550/arXiv.2103.14030

Milletari, F., Navab, N., and Ahmadi, S. A. (2016). "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)* (IEEE). doi: 10.48550/arXiv.1606.04797

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., et al. (2018). "Attention U-Net: Learning Where to Look for the Pancreas." *arXiv* abs/1804.03999. doi: 10.48550/arXiv.1804.03999

Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., and Shlens, J. (2019). *Stand-alone self-attention in vision models* (NIPS). doi: 10.48550/arXiv.1906.05909

Ren, M., and Zemel, R. S. (2017). "End-to-end instance segmentation with recurrent attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6656–6664. doi: 10.1109/CVPR.2017.39

Romera-Paredes, B., and Torr, P. H. S. (2016). "Recurrent instance segmentation," in *European Conference on computer vision* (Springer), 312–329. doi: 10.48550/arXiv.1511.08250

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention*, 9351 234–241 . doi: 10.1007/978-3-319-24574-4_28

Scharr, H., Minervini, M., French, A. P., Klukas, C., Kramer, D. M., Liu, X., et al. (2016). Leaf segmentation in plant phenotyping: a collation study. *Mach. Vis. Appl.* 27 (4), 585–606. doi: 10.1007/s00138-015-0737-3

Shen, D., Wu, G., and Suk, H. I. (2017). Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* 19, 221. doi: 10.1146/annurev-bioeng-071516-044442

Singh, V. (2019). Sunflower leaf diseases detection using image segmentation based on particle swarm optimization. *Artif. Intel. Agric.* 3, 62–68. doi: 10.1016/j.aiia.2019.09.002

Slaughter, D. C., Giles, D., and Downey, D. (2008). Autonomous robotic weed control systems: A review. *Comp. Elec. Agric.* 61, 63–78. doi: 10.1016/j.compag.2007.05.008

Tian, K., Li, J., Zeng, J., Evans, A., and Zhang, L. (2019). Segmentation of tomato leaf images based on adaptive clustering number of K-means algorithm. *Comp. Elec. Agric.* 165, 104962. doi: 10.1016/j.compag.2019.104962

Ubbens, J. R., and Stavness, I. (2017). Deep plant phenomics: a deep learning platform for complex plant phenotyping tasks. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.01190

Wang, Z., Wang, K., Yang, F., Pan, S., and Han, Y. (2018). Image segmentation of overlapping leaves based on chan-vese model and sobel operator. *Inform. Proc. Agric.* 5 (1), 1–10. doi: 10.1016/j.inpa.2017.09.005

Ward, D., Moghadam, P., and Hudson, N. (2021). Deep leaf segmentation using synthetic data. *arXiv*. doi: 10.48550/arXiv.1807.10931

Xia, C., Lee, J. M., Li, Y., Song, Y. H., Chung, B. K., and Chon, T. S. (2013). Plant leaf detection using modified active shape models. *Biosyst. Eng.* 116, 23–35. doi: 10.1016/j.biosystemseng.2013.06.003

Zhang, K., Li, Y., Liang, J., Cao, J., Zhang, Y., Tang, H., et al. (2022). Practical blind denoising *via* swin-Conv-UNet and data synthesis. *arXiv*. doi: 10.48550/arXiv.2203.13278

Zhang, Z., Liu, Q., and Wang, Y. (2018). Road extraction by deep residual U-net. *IEEE Geosci. Remote Sens. Lett.* 15 (5), 749–753. doi: 10.1109/LGRS.2018.2802944

Zhou, Z., Siddiquee , M. M. R., Tajbakhsh , N., and Liang , J. (2019). UNet++: redesigning skip connections to exploit multiscale features in image segmentation . *IEEE Trans. Med. Imaging* 39 (6), 1856-1867. doi: 10.1109/TMI.2019.2959609