Engenharia Agrícola

ISSN: 1809-4430 (on-line)

www.engenhariaagricola.org.br

Scientific Paper
Doi: http://dx.doi.org/10.1590/1809-4430-Eng.Agric.v44e20230097/2024

SE-SWIN UNET FOR IMAGE SEGMENTATION OF MAJOR MAIZE FOLIAR DISEASES

Yujie Yang[1,2], Congsheng Wang[1,3], Qing Zhao[3,4], Guoqiang Li[3,4], Hecang Zang[3*,4]

[3*]Corresponding author. Institution of Agricultural Economy and Information, Henan Academy of Agricultural Sciences/Zhengzhou, 450002, China. E-mail: zanghecang@163.com | ORCID ID: https://orcid.org/0000-0002-5117-8744

KEYWORDS

maize leaf diseases; image segmentation; Swin-Unet; Swin transformer; SENet.

ABSTRACT

Maize yields are important for human food security, and the issue of how to quickly and accurately segment areas of maize disease is an important one in the field of smart agriculture. To address the problem of irregular and multi-area clustering of regions of maize leaf lesions, which can lead to inaccurate segmentation, this paper proposes an improved Swin-Unet model called squeeze-and-excitation Swin-Unet (SE-Swin Unet). Our model applies Swin Transformer modules and skip connection structures for global and local learning. At each skip connection, a SENet module is incorporated to focus on global target features through channel-wise attention, with the aims of highlighting significant regions of disease on maize leaves and suppressing irrelevant background areas. The improved loss function in SE-Swin Unet is based on a combination of the binary cross entropy and Dice loss functions, which form the semantic segmentation model. Compared to other traditional convolutional neural networks on the same dataset, SE-Swin Unet achieves higher mean results for the intersection over union, accuracy, and F1-score, with values of 84.61%, 92.98%, and 89.91%, respectively. The SE-Swin Unet model proposed in this paper is therefore better able to extract information on maize leaf disease, and can provide a reference for the realisation of the complex task of corn leaf disease segmentation.

INTRODUCTION

Food security is facing serious challenges due to the ongoing impact of the COVID-19 pandemic, the complex and changing environments at home and abroad, and the increasingly frequent occurrence of abnormal weather and natural disasters (Laborde et al., 2020). Experience with many production practices has shown that disease is an important factor affecting crop yield, as it can cause yield reductions of 10–40% or even crop failure, which can seriously threaten food security (Wu et al., 2021). The breeding of disease-resistant crop varieties for target cultivation environments using breeding techniques can effectively mitigate the damage caused by disease (Zeng et al., 2019; Mi et al., 2020). To breed resistance to maize disease, accurate and efficient disease phenotyping is required for variety screening, identification, and excellent genetic resource discovery. The maize industry in China is developing rapidly, with a total sown area of 41.26 million ha in 2020 and a high yield record of 24,948.75 kg/ha (Zeng, 2022). Maize diseases are increasing every year, and have become a significant factor affecting the yield and quality of maize. The primary leaf diseases of maize are known as large spot, small spot, and stripe rust, and these can significantly limit the normal growth of maize, leading to a decrease in yield and quality. Early detection of crop diseases can reduce economic losses and positively affect crop quality (Muhammed, 2022; Hussain et al., 2022). In general, disease mainly occurs on the leaves of plants, meaning that observation of the diseased areas of maize leaves can enable effective visual assessment of the type and extent of the disease site, thus aiding in the early diagnosis of disease and timely disease control. Image segmentation is a technique in which an image is divided into several specific regions with unique properties, and the target of

[1] School of Computer and Information Engineering, Henan Normal University/Xinxiang, 453007, China.
[2] Key Laboratory of Artificial Intelligence and Personalized Learning in Education of Henan Province/Xinxiang, 453000, Henan, China.
[3] Institution of Agricultural Economy and Information, Henan Academy of Agricultural Sciences/Zhengzhou, 450002, China.
[4] Huanghuaihai Key Laboratory of Intelligent Agricultural Technology, Ministry of Agriculture and Rural Affairs/Zhengzhou, 450002, China.

Area Editor: Gizele Ingrid Gadotti
Received in: 7-3-2023
Accepted in: 12-15-2023

interest is extracted. It is a key step in image processing for image analysis, and segmentation accuracy of leaf disease regions directly affects the accuracy of disease identification. Improving the accuracy of leaf image segmentation to provide farmers with optimal solutions for managing crop diseases has therefore become a focus of current research.

Traditionally, the disease condition of maize leaves has been judged manually, which is not only an inefficient and error-prone approach, but is also heavily reliant on personal experience. However, in recent years, with the development of computer vision technology for agriculture, there have been many advancements in crop disease segmentation research. Common traditional image segmentation algorithms used in agriculture include methods based on thresholding (Wang & Guo, 2018; Wang et al., 2018), clustering (Guo & Li, 2015; Huo et al., 2019), region growing (Xu et al., 2017; Zhang et al., 2020), and graph theory (Shi & Malik, 2000; Cong et al., 2018). Although these methods are easy to implement and operate, they require manual extraction of image features, use a single approach, and are difficult to generalise for image segmentation. Following the rapid developments in deep learning, however, image segmentation methods based on deep learning have become a research hotspot. These methods extract the pixel-level features of leaf image spots and carry out semantic segmentation via self-learning. Compared to traditional methods, algorithms based on deep learning save a lot of work and time for humans, and offer superior performance to traditional methods. Hence, numerous semantic segmentation methods based on deep learning have been introduced into the field of agricultural image segmentation, such as the semantic segmentation of cotton canopy images using fully convolutional networks (FCNs) and conditional random field (CRF) network models (Shelhamer et al., 2017; Liu et al., 2018); an improved FCN network applied to solve the segmentation of corn leaf spots (Wang et al., 2019); a U-net structure used to solve the semantic segmentation problem of maize leaf disease images (Liu et al., 2021a); and an improved DeepLabv3+ deep learning network for segmentation of black rot spots in grape leaves (Yuan et al., 2022). Wiesner-Hanks et al. (2019) used images taken by unmanned aerial vehicles (UAVs) as training data and a convolutional neural network (CNN) for deep learning, and then passed the output to a CRF for post-processing, making it possible to achieve accurate segmentation of diseased and non-diseased areas of maize leaf images. This method was capable of detecting millimetre-scale plant diseases through the use of deep learning and crowdsourced data. Huang et al. (2021) proposed an image segmentation method based on YOLACT++ and an attention module for segmenting lesions in maize leaves under natural conditions, with the aim of improving the accuracy and real-time performance of lesion segmentation. However, these traditional CNNs still suffer from issues such as under-segmentation and low segmentation efficiency. A deep network model loses some spatial contextual information in the process of down-coding to extract higher-level semantic features, which affects the subsequent segmentation accuracy (Wang et al., 2020).

The integration of attention strategies with deep learning has made the task of identifying and segmenting plant disease regions more detailed and attractive. Some major attention modules that are commonly used include the convolutional block attention module (CBAM), squeeze-and-excitation network (SENet), and visual-spatial-graph network (VSG-Net) (Woo et al., 2018; Hu et al., 2018; Ulutan et al., 2020). There have recently been breakthroughs in the field of computer vision (CV) due to the good performance of Transformers, and many new Transformer-based methods for CV tasks have been proposed (Dai et al., 2021). Swin-Unet is a Transformer-based segmentation network that performs well on computed tomography (CT) images of the liver (Cao et al., 2021). This model is mainly used for medical image segmentation tasks, and researchers have not yet improved the model to solve problems in agriculture.

The diversity of symptoms related to maize disease leads to irregular and multiregional clustering of maize leaf lesion regions, making it really challenging for segmentation models to identify and localise these irregular regions. To solve these problems, this paper proposes an improved Swin-Unet model called squeeze-and-excitation Swin-Unet (SE-Swin Unet). In SE-Swin Unet, a preprocessing method is applied to the image, and an image enhancement method is designed to improve its performance. By combining the loss functions and evaluation indexes used for typical binary classification in the segmentation field, various loss functions are designed. Finally, a hybrid loss function is constructed by comparing the performance of these different loss functions on the test set, based on a combination of the binary cross-entropy (BCE) loss, and an SENet module is added at the skip connections to form the SE-Swin Unet semantic segmentation model. The features of diseased regions in images of maize leaves are obtained by training, and disease spot segmentation is performed for several common types of maize leaf diseases. At the end of this study, several comparative experiments are carried out to verify the effectiveness of the proposed method.

MATERIAL AND METHODS

In this section, we first describe the preprocessing of the dataset. Next, we add the SENet module to Swin-Unet and improve the loss function to develop a new model called SE-Swin Unet. We then introduce a contrastive network based on a CNN. Finally, experiments are conducted in a specific environment, and appropriate evaluation metrics are used to assess the effectiveness of the model.

Dataset

In this study, the PlantVillage dataset (Hughes & Salathé, 2015) was used as the source of training data. This dataset contains 54,303 images of healthy and diseased plants, which are categorised into 38 classes including maize, apple, blueberry, grape, peach, potato, tomato, and strawberry. Specifically, 800 images of maize disease were selected from the dataset for this study, including 300 images of maize small spot disease, 200 images of maize stripe rust disease, and 300 images of maize large spot disease. Examples of these images are shown in Figure 1.

FIGURE 1. Images of three maize leaf diseases: (a) large spot disease; (b) small spot disease; (c) stripe rust.

Images of single leaves were collected that contained irregular areas of maize leaf disease, with an image size of 256 × 256 pixels. The images were saved in .jpg format to construct a sample set of images of maize leaf disease. To train and evaluate the network model, the data were divided into a training set (80%), validation set (10%), and test set (10%).

Image preprocessing

As the PlantVillage dataset lacked the image labels that would be required for supervised training of computerised deep learning, the preprocessing step involved manual labelling of the 800 images using the open-source labelling tool Label ME (Russell et al., 2008) from the Massachusetts Institute of Technology (MIT). The diseased area of the corn leaves in the mask image was marked with a pixel value of 255, and the rest of the image was labelled as the background area with a pixel value of zero. Each labelled image was saved in .png format. Examples of maize leaf disease images from the dataset with their mask annotations are shown in Figure 2. As the dataset contained only 800 annotated maize leaf disease images, we applied data augmentation techniques to the manually annotated images to improve the generalisation ability of the network model during the training process.

To improve the robustness of the model, we employed an online data augmentation method, which is a technique drawn from modern deep learning frameworks. In this approach, before each training epoch, the dataset was augmented using techniques such as flipping, rotating, panning, and scaling, In addition, features extracted from the hue, saturation and value (HSV) colour space transformation give better results. In each transformation method, a random factor was applied to ensure that the data were unique for each training epoch, which helped to enhance the diversity of the training data. More specifically, the images and labels underwent simultaneous random flipping and rotation within the range [−20°, 20°]. In the HSV domain, the hue channel was varied randomly within the offset range [−180, 180], while the offsets for the saturation and value channels were within the range [−255, 255]. The input image also underwent random affine transforms, such as panning, scaling and rotation, to provide additional variety during the training process.

FIGURE 2. Interface used to mark the locations of maize leaf disease in an image.

SE-Swin Unet Method

In this paper, we build an SE-Swin Unet model by introducing the SENet module into Swin-Unet and improving the loss function.

SENet

SENet enhances the accuracy by modelling the correlation between feature channels and reinforcing important features (Hu et al., 2018), using the architecture illustrated in Figure 3. SENet comprises two essential parts: the squeeze operation, and the excitation operation. Traditional convolution can only operate in a local space, making it difficult to extract the relationship between individual channels. Hence, SENet uses the squeeze operation to obtain sufficient information. The specific operation for each channel is shown in [eq. (1)].

$$z_c = f_{sp}(u_c) = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_{c(i,j)} \quad (1)$$

where:

$f_{sp}(u_c)$ is the squeeze operation on a matrix u_c;

H is the height of the matrix, and

W is the width of the matrix.

The squeeze operation captures the global features that correspond to all channels, while the excitation operation is designed to model inter-channel relationships. In the excitation step, each channel learns the activation of specific samples through a channel-dependent, self-gating mechanism. It learns to use global information, to selectively focus on informative features, and to suppress less useful information characteristics. A sigmoid function is used to introduce nonlinearity, and a ReLU activation function is inserted in the middle to limit the complexity of the model and improve the training process, as shown in [eq. (2)]:

$$s = F_{ex}(z,W) = \sigma(g(z,W)) = \sigma\left(W_2 \delta(W_1 z)\right), W_1 \in R^{\frac{C}{r}}, W_2 \in R^{C \cdot \frac{C}{r}} \quad (2)$$

where:

z is the result obtained of the squeeze operation;

σ is the sigmoid function;

δ is the ReLU function,

W_1 , W_2 are the fully connected functions for compression and reconstruction, respectively.

The normalised weights learned from the excitation operation are applied to the features of each channel, which are then combined to produce the final output of the block,

$$X_c' = F_{scale}(u_c, s_c) = S_c u_c \quad (3)$$

where:

F_{scale} is defined for the reweighting operation,

S_c is the output matrix channel of the excitation operation.

As can be seen from Figure 3, the SENet module first performs a squeeze operation on the feature map, using a global average pooling operation in the channel dimensional direction to make the receptive field wider. An excitation operation is then performed on the global features, which is completed by two fully connected layers plus the ReLu activation function, which are used to build a complex feature representation to learn the relationships between the channels. The weights of the different channels are then obtained by the sigmoid activation function, and are finally output with the original feature channels weighted by the scale layer. Essentially, the purpose of the SE module is to make the network focus more on the most useful feature channels while suppressing the useless ones, to achieve feature representation and generalisation more effectively.

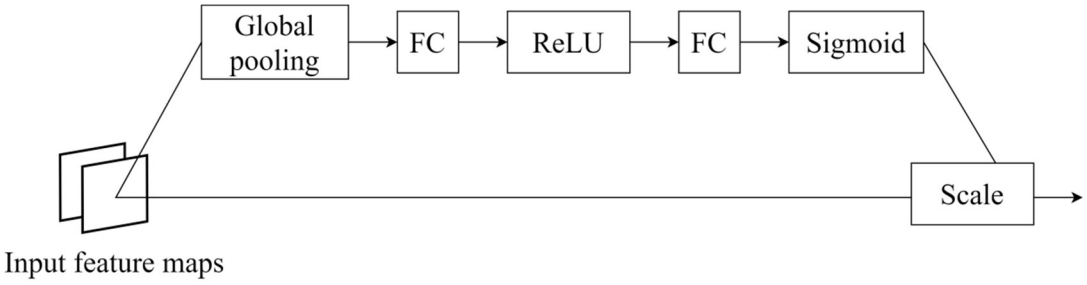

FIGURE 3. Structure of the SENet module.

Architecture of SE-Swin Unet

In the SE-Swin Unet model, we introduce the SENet module at each skip connection, with the aim of weighting the feature maps through the channel attention mechanism, thus improving the model's ability to capture details and spatial information and further enhancing the segmentation performance of the model. In each skip connection, we first perform global average pooling of the input feature maps using a global pooling layer to obtain the global importance of each channel. The global description is then converted into a vector with the same number of input channels through a layer of fully connected networks. This vector is processed by an activation function such as sigmoid to obtain the weights of each channel. Multiplying these weights by the original feature map forms the weighting operation for the different channel information.

Finally, we take the weighted features as output to enhance the original feature representation. In this way, the SENet module enables the model to focus on the features

that are more important for the task at hand by dynamically adjusting the importance of the channels.

In the proposed SE-Swin Unet, at each skip connection, the feature tensor of the skip connection is first taken as the input to the SENet module. The squeeze and excitation operations are then applied to obtain the feature weighting results. Finally, the output of the SENet module is added to the corresponding feature tensor of the decoder, which forms the output feature tensor at the skip connection for subsequent decoder operations. This approach enhances the connection between low-level and high-level features, resulting in finer features for multi-scale prediction and segmentation. Due to the irregularity of the regions of maize leaf disease, the SENet attention mechanism is used to quickly select information and allocate more resources to the target region for more accurate segmentation. The SENet attention mechanism is introduced at each skip connection to enable the model to focus on significant regions of maize leaf disease while suppressing irrelevant background regions. The architecture of SE-Swin Unet is shown in Figure 4, and consists of an encoder (the left-hand side of Figure 4), a bottleneck, a decoder (the right-hand side of Figure 4), and skip connections (the middle span section of Figure 4), where each component is composed of Swin Transformer blocks (Liu et al., 2021b). Detailed descriptions of the sections of SE-Swin Unet are given below.

Encoder: To enable the sequence input, the patch partition splits the input image into non-overlapping patches of size 4×4. Each patch is then transformed into a tensor format that can be processed by the Swin Transformer using a linear embedding layer. In the patch merging layer, the patches that are divided into four parts are joined together, resulting in a halving of the patch resolution. The merged feature dimension is then four times the size of the original, so a linear layer is added to the concatenated features to unify the feature dimension back to twice the original size. Specifically, the patch merging layer is responsible for the downsampling operation, while the Swin Transformer block is used to learn the feature representation. This process needs to be repeated three times.

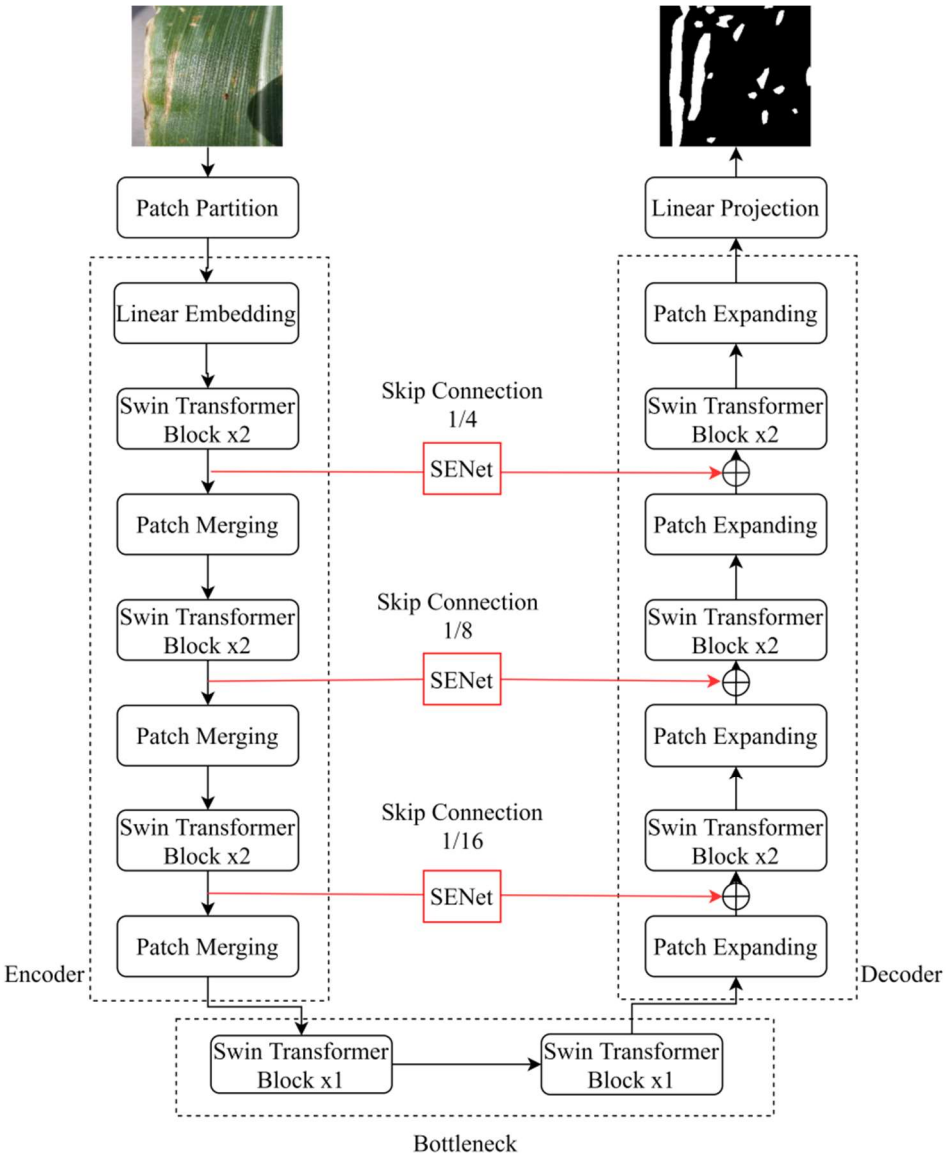

FIGURE 4. Structure of SE-Swin Unet.

Loss Function

The loss function is essential in assessing the quality of network learning, as it quantifies the discrepancy between the predicted values from the model and the actual ground truth values (Chen et al., 2016). Given the diverse range of image segmentation tasks, the Dice loss function introduced in VNet (Yang et al., 2019) was specifically designed to address situations where there is a significant imbalance between the numbers of positive and negative samples in a semantic segmentation process. In the case of corn disease images, the diseased regions exhibit varying sizes and shapes, and some entire corn leaves may be marked as disease regions. The Dice loss function is more suitable for highly imbalanced samples. However, using this loss function can pose limitations in terms of backpropagation, making it prone to gradient explosion and leading to unsatisfactory segmentation outcomes. In contrast, the cross-entropy (CE) loss is used in multiple classification tasks. To achieve better segmentation performance and determine the most appropriate loss function, we combine the Dice loss function with the binary classification BCE loss function to give the loss function in [eq. (8)]:

$$\text{loss} = 0.4 * \text{loss Dice} + 0.6 * \text{loss BCE} \quad (4)$$

Bottleneck: Due to the depth of the Transformer, convergence is not achievable in the bottleneck. Hence, only two consecutive Swin Transformer blocks are used to construct a bottleneck for learning deep feature representation. At this stage, the feature dimension and resolution remain unchanged.

Decoder: At this stage, unlike the patch merging layer used in the encoder, we use a patch expansion layer to apply upsampling operations in order to achieve image reconstruction of adjacent dimensional feature maps. This process doubles the resolution of the image features while reducing the corresponding feature dimension to half of the original size. In fact, the purpose of upsampling the image features is achieved in the patch expanding layer, and the dimensionality of the input features is first doubled via a linear layer. A rearrangement operation is then used to double the resolution of the input features while reducing the feature dimension to one-fourth of the size before the rearrangement operation.

Skip connections: At each skip connection, we incorporate an SENet module, which combines the multi-scale features from the encoder with the upsampled features. This fusion integrates shallow and deep features to alleviate the loss of spatial information caused by downsampling. A linear layer is also added to maintain the same dimensions between the connected features and the upsampled features. In particular, the SENet module captures the low-level features before the skip connections and the high-level features from the decoder output. It then generates attention weights to combine low-level features from different channels, resulting in accurate feature representations.

Swin Transformer block: The encoder, bottleneck and decoder are all constructed based on the Swin Transformer block. Unlike the traditional multi-headed self-attentive (MSA) module, the Swin Transformer block is based on a shifted window mechanism. The proposed Swin Transformer block is composed of a layer norm (LN) layer, an MSA module, a residual connection, and a two-layer multilayer perceptron (MLP) with GeLu nonlinearity, as shown in Figure 5. Two consecutive Swin Transformer blocks are used: the first block uses a window-based multi-headed self-attentive (W-MSA) module, while the second uses a displaced window-based multi-headed self-attentive (SW-MSA) module. Based on this window partitioning mechanism, the consecutive Swin Transformer blocks can be expressed as shown in eqs (5)–(8).

$$\hat{z}^{l} = \text{W-MSA}\left(\text{LN}\left(z^{l-1}\right)\right) + z^{l-1} \quad (5)$$

$$z^{l} = \text{MLP}\left(\text{LN}\left(\hat{z}^{l}\right)\right) + \hat{z}^{l} \quad (6)$$

$$\hat{z}^{l+1} = \text{SW-MSA}\left(\text{LN}\left(z^{l}\right)\right) + z^{l} \quad (7)$$

$$z^{l+1} = \text{MLP}\left(\text{LN}\left(\hat{z}^{l+1}\right)\right) \quad (8)$$

where:

$\hat{z}^{l}$ is the output of the SW-MSA module,

z^{l} is the output of the MLP module of the lth block.

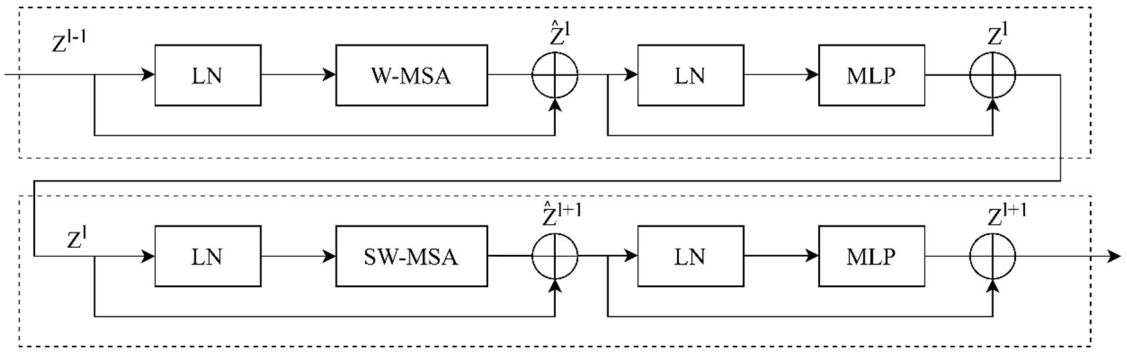

FIGURE 5. Structure of the Swin Transformer block.

In this hybrid loss function, the weights of the Dice and BCE loss functions are 0.4 and 0.6, respectively, and are determined based on the results of binary classification and empirical optimisation of maize leaf disease data in this paper. The weight for the Dice loss is lower, and it has a relatively small effect on the optimisation of the model, while the weight of BCE loss is set to 0.6, meaning that more attention is paid to the performance evaluation of the binary classification task when calculating the overall loss. The latter therefore plays a greater role in the optimisation of the model, thus improving the segmentation effect.

CNN-based Network

This paper compares two CNN-based segmentation models using two backbone networks to achieve different segmentation effects: U-Net (Ronneberger et al., 2015) and DeeplabV3+ (Chen et al., 2018). U-Net is an improved version of the FCN that uses a combination of up-sampling and down-sampling to achieve pixel-level image segmentation while minimising the loss of semantic information. U-Net is well-suited to small training sets, and has a U-shaped, left-right symmetric architecture. It uses two different backbone networks, Visual Geometry Group (VGG) and ResNet50. Multiple 3 × 3 convolutional kernels are used rather than large kernels, to reduce the number of required parameters, and 2 × 2 pooling kernels are used for maximum pooling. All of the hidden layers use the rectified linear unit (ReLU) activation function. ResNet50 is a deep residual network in which the problem of gradient disappearance during the training period is solved by allowing network layers to perform equal mapping. It has four large groups of 3, 4, and 6, and three small blocks, each with three convolutions.

DeepLabV3+ is another encoder-decoder model that achieves more accurate boundaries by using a spatial pyramid pooling technique to combine multi-scale convolutional features from multiple models. It uses two different backbone networks, MobileNet and Xception. MobileNet is a lightweight deep neural network that was proposed by Google in 2017, in which depthwise separable convolutions form the basic building blocks. Compared to traditional convolutional networks, MobileNet significantly reduces the number of parameters and computations while maintaining accuracy. Xception is an improved version of Inception that uses depth-separable convolution to decouple the cross-channel and spatial correlations to some extent.

Evaluation Metrics

To ensure the effectiveness of our model after training, we use three evaluation metrics, the accuracy, mean intersection over union (MIOU), and F1-score, to evaluate and compare the accuracy of the maize leaf disease image segmentation results, where higher quantitative values indicate better segmentation performance. The accuracy is used to determine the proportion of correct classifications, whereas the MIOU is a standard measure of semantic segmentation, and is calculated based on the average of the ratio of intersection and union of all categories. The F1-score is the summed average of the precision and recall. The formulae used for these calculations are shown in eqs (9)–(11).

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

$$\text{MIOU} = \frac{TP}{FP+FN+TP} \tag{10}$$

$$\text{F1-score} = \frac{2\times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

where:

TP is the model correctly detected as a positive case, representing true positive.

FP is the model error detected as a positive case, representing false positive.

TN is the model correctly detected as a negative example, representing true negative.

FN is a negative example of model error detection, representing false negative.

Experimental Environment

Table 1 presents the details of the hyperparameters used in this paper. In the experiments, we used version 1.1.0 of the Pytorch deep learning framework, running on a 7-core Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz with an RTX 2080 Ti GPU with 11 GB video memory. We used version 10.0 of Cuda, and version 3.7 of Python. To ensure the validity and accuracy of the experimental data, all experiments in this paper were conducted on the same computer. The model was trained for 100 epochs, with a batch size of four, an initial learning rate of 0.0001, and a momentum parameter of 0.5.

TABLE 1. Hyperparameter settings used for training.

Name	Value	Description
Epochs	100	Number of times the model was trained
Batch size	4	Number of samples selected for one training
Momentum	0.5	Used to prevent model overfitting
Learning rate	0.0001	Tuning parameter for optimisation algorithms
Input size	256 × 256	Size of the image input into the model

RESULTS AND DISCUSSION

In this section, to find the optimum loss function, we explore and evaluate different loss functions for training of the model, and finally develop a hybrid BCE+Dice loss function. To improve the attention and importance assigned by the neural network to the input data, we discuss the potential use of the SE module, which improves the performance of the model through an attention mechanism. Finally, we report comparative experiments with other networks and analyse the results.

Comparison of Loss Functions

Since different loss functions have different characteristics and suitability to various tasks and models, we conduct a comparative experiment with hybrid CE+Dice and BCE+Dice loss functions to evaluate the segmentation performance. Table 2 shows the segmentation results obtained with these two different loss functions. The results demonstrate that the BCE+Dice hybrid loss function yields better segmentation performance; this is because the BCE loss function guides the Dice loss function, and the BCE+Dice hybrid loss function combines the strengths of both. This approach allows for directional refinement during network training backpropagation, especially when dealing with complex samples that are difficult to learn. This leads to more stable learning, alleviates the problem of category imbalance, and improves the segmentation performance of the model. Hence, in this paper, we choose the BCE+Dice hybrid loss function.

TABLE 2. Comparison of loss functions.

Method	Loss function	MIOU(%)	Accuracy(%)	F1-Score(%)
Swin Unet	CE+Dice	80.04	88.96	82.40
Swin Unet	BCE+Dice	84.13	91.54	88.51

Figure 6 shows the curves of the loss values over the training epochs for the two hybrid loss functions. From Figure 6, we see that when the number of epochs exceeds 70, the loss value of CE+Dice stabilises and decreases to 0.117, and the loss value of BCE+Dice stabilises at 0.057, with fluctuations within 0.01. The trends in the behaviour of the training loss and validation loss curves are consistent, indicating that the proposed model performs similarly on the training and validation sets. In this case, we can conclude that the model has good generalisation ability, can adapt to new data, and has a high level of robustness.

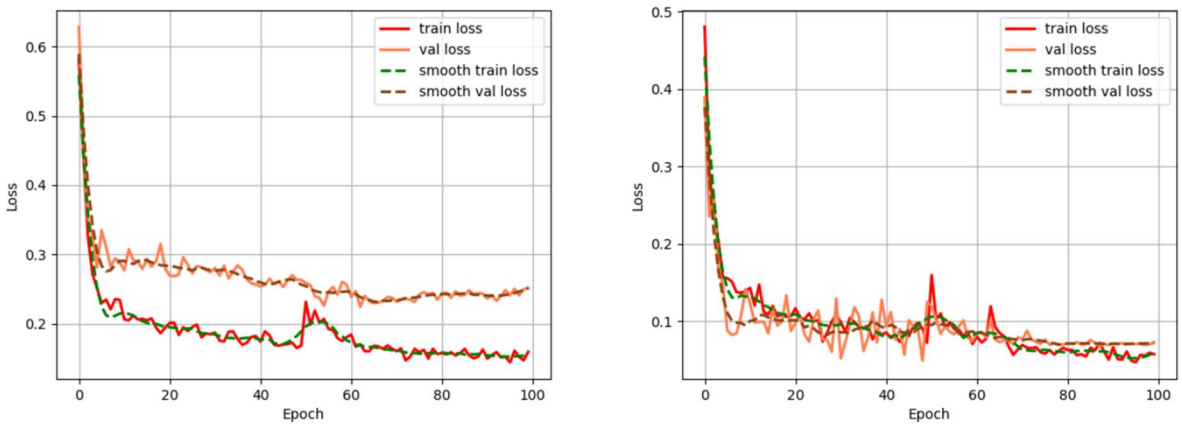

FIGURE 6. Curves showing the change in the hybrid loss functions over the training epochs: (a) CE+Dice; (b) BCE+Dice.

Evaluating the Impact of the SENet Module on Model Performance

SE blocks are used to adaptively recalibrate channel feature responses, which allows the network to selectively emphasise informative features and to suppress less useful ones using global information (Tan & Xiang, 2022). An attention mechanism is employed to quickly select information and allocate more resources to the target region, resulting in more accurate segmentation. We therefore introduce an attention mechanism at each skip connection to enable the model to focus on significant regions of maize leaf disease while suppressing irrelevant background regions. The results demonstrate that the addition of the SENet module improves the segmentation performance of the Swin-Unet model for corn leaf disease images. Table 3 presents the segmentation results with and without the SENet module, where 'N' means without the SENet module and 'Y' means with the SENet module. It is clear that the model with the SENet module has better performance, and we therefore incorporate SENet module into our model.

TABLE 3. Comparison of results with and without the SENet module.

Method	SENet block	MIOU(%)	Accuracy(%)	F1-Score(%)
Swin Unet	N	84.13	91.54	88.51
Swin Unet	Y	**84.61**	**92.98**	**89.91**

Comparison with CNN-based Networks

After the training process, several classical CNN-based networks were selected in order to evaluate their results in comparison with SE-Swin Unet on the test set. The evaluation metrics applied to each model for the segmentation of corn leaf disease images are presented in Table 4. From the table, we see that SE-Swin Unet outperforms all other CNN-based models on each evaluation metric.

TABLE 4. Comparison of segmentation performance of different models.

Model	Backbone	MIOU(%)	Accuracy(%)	F1-Score(%)
U-Net	VGG	83.30	90.97	87.13
	ResNet50	83.25	90.93	87.04
DeeplabV3+	MobileNet	80.91	89.16	86.52
	Xception	79.69	88.85	85.31
Swin Unet	Swin Transformer	80.04	88.96	82.40
SE-Swin Unet	Swin Transformer	**84.61**	**92.98**	**89.91**

When applied to the task of segmentation of corn leaf disease images, the performance of the models based on the U-Net and DeeplabV3+ frameworks is similar, but the choice of backbone network also impacts the accuracy of the results. For the DeeplabV3+ model, the use of the MobileNet backbone slightly improves the MIOU and F1-score by 1.22% and 1.21%, respectively, compared to Xception. For the U-Net model, VGG has better overall performance than ResNet50, whereas the performance of SE-Swin Unet is superior to that of VGG and both of the DeeplabV3+ models. In terms of MIOU, SE-Swin Unet outperforms Xception by 4.92%, and in terms of accuracy, it outperforms all other models by more than 2%. SE-Swin Unet has the highest metrics for MIOU, accuracy and F1-score, with values of 84.61%, 92.98%, and 89.91%, respectively. These results demonstrate that SE-Swin Unet has good adaptability and high accuracy when applied to images of maize leaf disease.

Table 4 shows that the lighter backbone networks perform better for corn leaf disease images, as VGG outperforms ResNet50, and MobileNet outperforms Xception. The deeper structured ResNet50 and Xception backbone networks lead to poor performance, due to slower fitting and fluctuating accuracy. The Swin Transformer uses a global self-attentive mechanism in which skip connections are added to the SENet module to enable the model to focus on the global target features. The experimental results show that the SE-Swin Unet model proposed in this paper achieves higher performance.

TABLE 5. Comparison of segmentation parameters for different methods for corn leaf disease images.

Method	Backbone	Flops(G)	Params(M)	Time(ms)
U-Net	VGG	17.61	24.89	7.1
	ResNet50	43.25	43.93	10.5
DeeplabV3+	MobileNet	5.04	5.81	7.2
	Xception	6.34	6.85	7.3
Swin Unet	Swin Transformer	88.05	120.96	19.7
SE-Swin Unet	Swin Transformer	94.12	122.58	28.4

Table 5 shows the results for the segmentation parameters of different methods for the segmentation of images of corn leaf disease. It can be seen that our network model has the largest number of parameters of all the network models compared here, with 94.12 G flops and 122.58 M parameters. This large number of parameters is mainly because our network model, SE-Swin Unet, adopts the structure of the Swin Transformer, and designs its own data enhancement techniques to improve the robustness and generalisation ability of the model. Our network also incorporates the SENet attention mechanism, which leads to an increase in the number of parameters and slower processing speed (with an average segmentation time of 28.4 ms). Our model has a stronger learning ability and expressive power, giving better segmentation detail, and therefore shows better performance on the complex task of corn leaf disease segmentation.

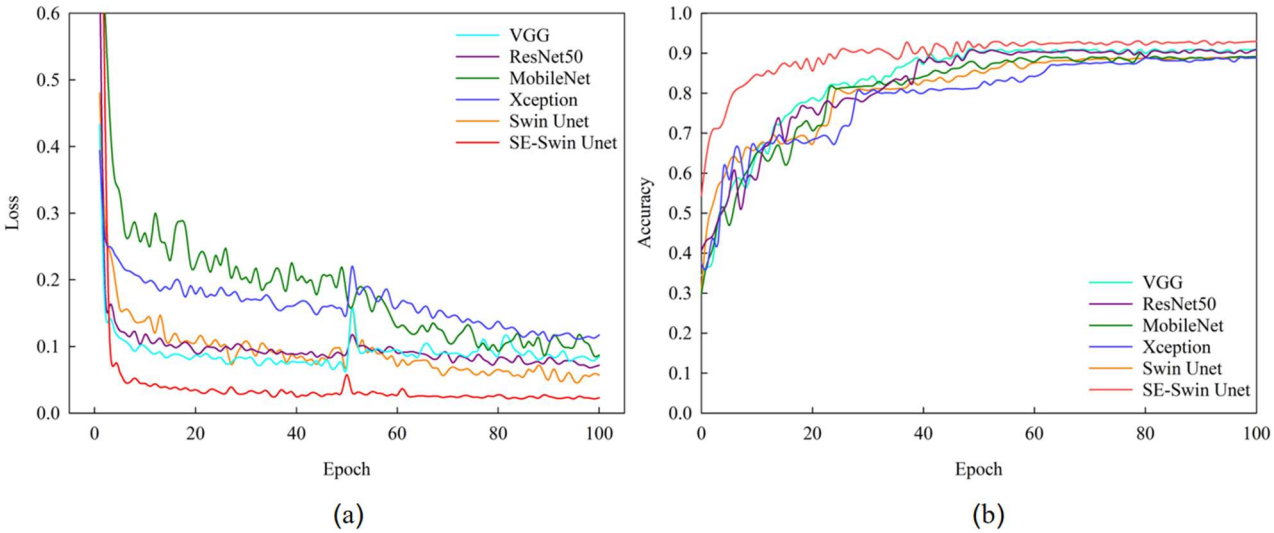

FIGURE 7. Comparison of loss and accuracy during the training of different networks: (a) change in loss; (b) change in accuracy.

Figure 7 shows the changes in the loss and accuracy for different models over the training epochs. From Figure 7(a), it can be seen that the loss values for all six network models have dropped rapidly by the 10th epoch, and are close to stabilising after the 80th epoch. SE-Swin Unet outperforms the other five models, and achieves the best convergence performance after the 60th epoch. From Figure 7(b), we see that the accuracy of each of the six network models gradually increases up to the 40th epoch, although the accuracy of our model rises and stabilises the most rapidly, meaning that our SE-Swin Unet model has better performance throughout the training process.

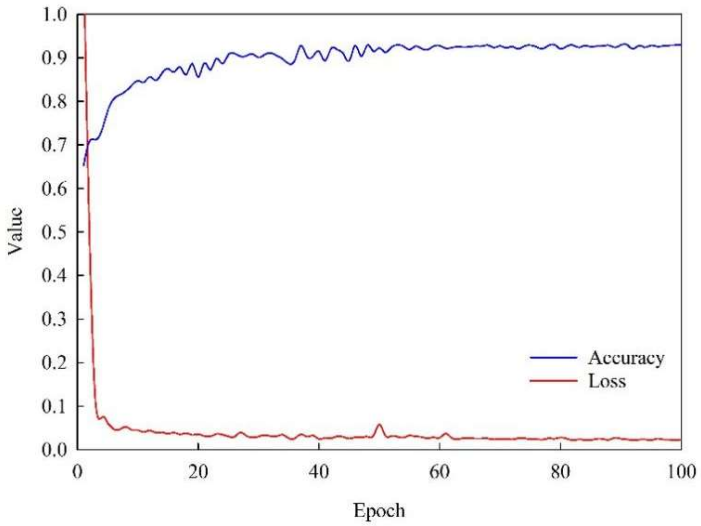

FIGURE 8. Loss and accuracy change curves for the SE-Swin Unet model.

Figure 8 shows the changes in loss and accuracy for the SE-Swin Unet model. It can be seen that as the number of training epochs increases, the accuracy of SE-Swin Unet increases and the loss value decreases. In particular, when the number of epochs reaches 50, the accuracy rate is stable at 0.92, the floating change is stable within 1%, the loss value is stable at 0.117, and the up and down fluctuations are within 0.01, indicating that the accuracy of the model is high and the robustness is good. The loss and accuracy curves indicate that the model proposed in this paper has strong fitting and generalisation capabilities. The experimental results show that the SE-Swin Unet model has superior segmentation performance.

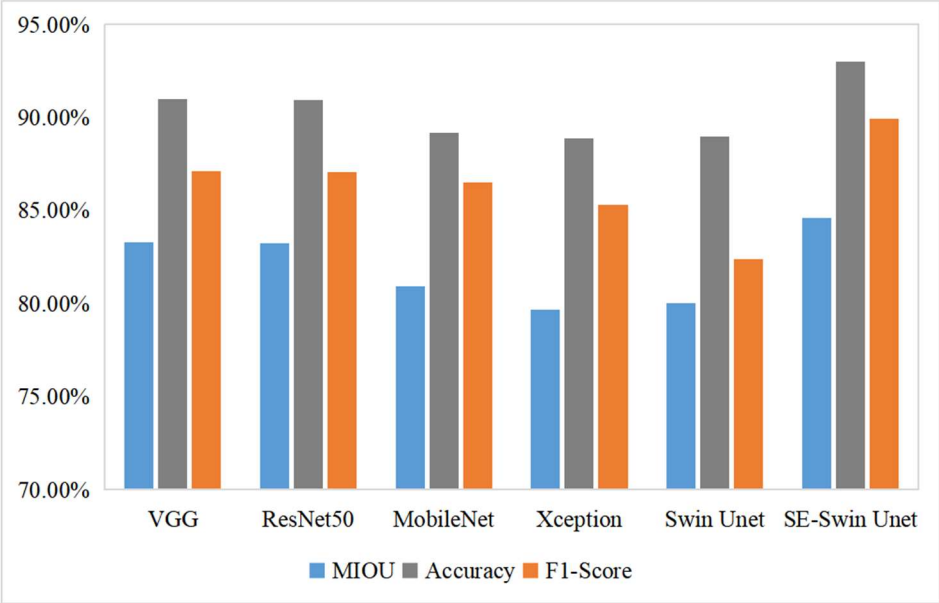

FIGURE 9. Differences in evaluation indicators for various networks.

Figure 9 displays histograms of the results obtained from various deep learning-based network models for measuring corn leaf disease images. It is evident that the accuracy achieved by the MobileNet and Xception models is below 90%, with Xception giving the lowest accuracy of 88.85%. This suggests that the relatively complex structure of the Xception network may reduce its generalisation ability. The accuracy values for the other models are above 90%, with SE-Swin Unet achieving the highest accuracy of 92.98%. These results indicate that the addition of the SENet module enhances the basic feature extraction ability of the network. SE-Swin Unet yields superior performance in the segmentation of corn leaf disease images compared to the other CNN-based models, and achieves a higher MIOU, accuracy, and F1-score, with increases in the ranges 1.31–4.92%, 2.01–4.13%, and 2.78–4.60%, respectively. The MIOU value for the proposed model is 84.61%, which is the highest among the models considered here. These results confirm the effectiveness of SE-Swin Unet in accurately segmenting regions of corn leaf disease.

Figure 10 presents some examples of results of corn leaf disease segmentation for some test sets. Each set, from left to right, shows the original image, the labelled image, the output of the proposed method, and the output of the other CNN-based methods for corn leaf disease images, respectively. It can be see that the different segmentation models perform similarly in segmenting maize leaves with large disease-covered areas in rows E and F, resulting in good segmentation. However, the CNN-based models exhibit under-segmentation and mis-segmentation for images with irregular areas of disease, possibly due to the localisation of the convolution operation. Our SE-Swin Unet method produces clearer and more complete contours, leading to better overall segmentation performance compared to the other methods.

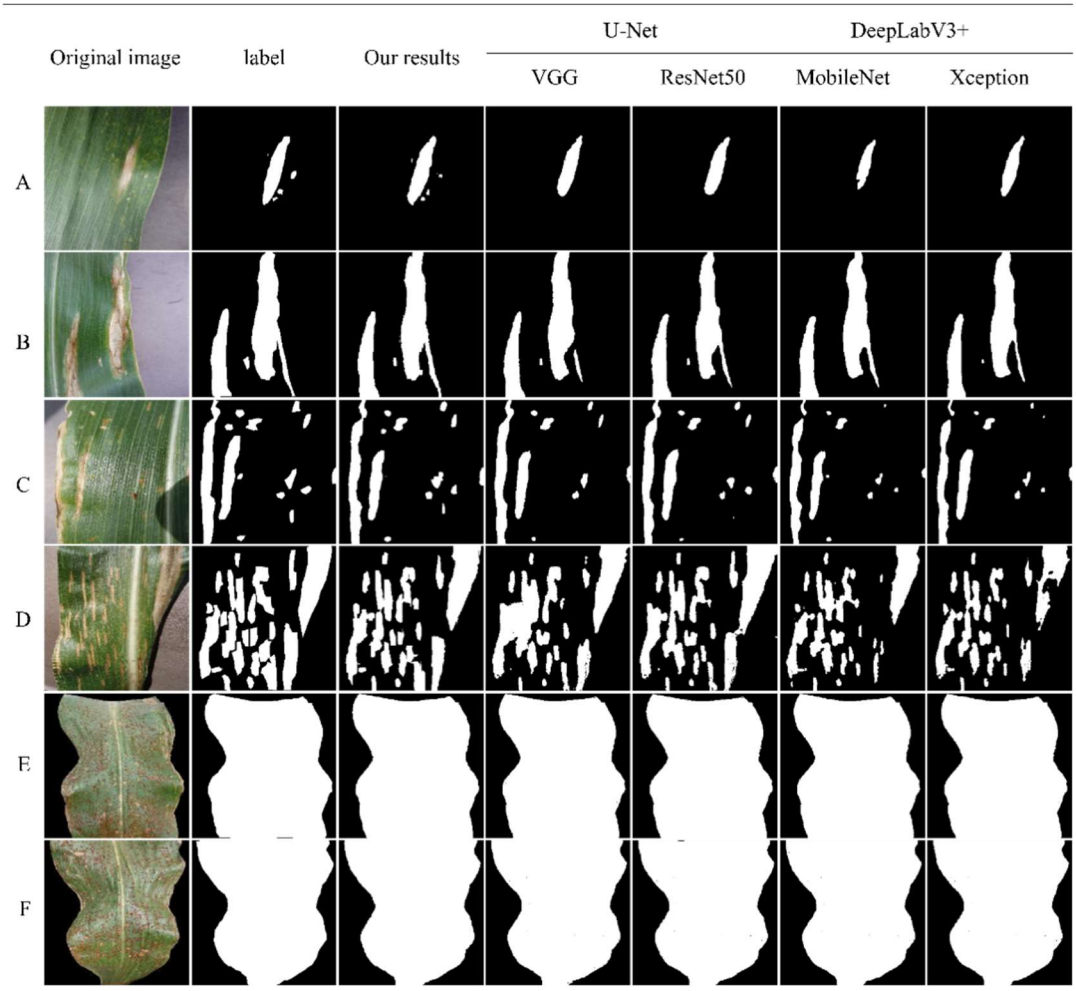

FIGURE 10. Comparison of different methods for the segmentation of maize foliar disease.

Discussion

Khan et al. (2023) proposed a weed-crop segmentation method for UAV images using a coder-decoder architecture and incorporating a dense spatial pyramid pooling module that was capable of extracting multi-scale features and seamlessly capturing local and global contextual information. This model achieved an MIOU of 0.81 on a challenging dataset, thus proving its feasibility. The attention module was a valuable component, as it helped the model to better focus on the important parts of the image, thus improving the accuracy of segmentation.

The tasks of maize leaf disease segmentation and weed segmentation have some commonalities in terms of the datasets, image features and processing methods used. Based on these commonalities, we can draw on our research results and experience of maize disease leaf segmentation to explore the weed segmentation problem in future work, in order to improve the generalisation ability and scope of application of our model.

From the experimental results in Table 5, it can be seen that the main shortcoming is that the number of parameters for SE-Swin Unet is larger than for the traditional U-Net and DeeplabV3+ models. In future, model compression and pruning techniques could be applied to reduce the number of parameters and to improve the efficiency and interpretability of the model.

In summary, compared with the other deep learning semantic segmentation models considered in this paper, the proposed method is more suitable for maize leaf disease segmentation, although there is still room for improvement. The next step will be to carry out more comparative tests for the above problems to make the model more accurate and efficient.

CONCLUSIONS

This paper has presented an improved Swin-Unet learning model that achieves good performance in terms of accurately segmenting maize leaf disease regions. To address the challenge of irregular and multi-area clustering of corn leaf disease regions, which can result in inaccurate segmentation, the SENet module is included in the skip connection process, and a hybrid loss function is used to optimise the model so that it can better help the model learn samples. Compared to other CNN-based models, such as the U-Net model with VGG and ResNet50 as backbone networks, and the DeepLabV3+ model with MobileNet and Xception as backbone networks, the proposed approach leads to significant improvements in segmentation accuracy, and is effective in accurately identifying and segmenting corn leaf disease regions. However, although the proposed network represents some progress in terms of segmentation accuracy and segmentation performance, there is still room for improvement in terms of reducing the number of model parameters to increase the efficiency of the model.

FUNDING

This paper was partly funded by the Key Research Projects of Henan Higher Education Institutions (24B120004), partly funded by the PhD Research Startup Support Project of Henan Normal University (QD19013), partly funded by key science and technology projects in Henan Province (212102110253), and partly funded by the Key Research and Development Program of China (2022YFD2001005).

REFERENCES

Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, Wang M (2021) Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv: 2105.05537.

Chen L, Qu H, Zhao J, Chen, B, Principe JC (2016) Efficient and robust deep learning with correntropy-induced loss function. Neural Computing and Applications 27: 1019-1031.

Chen L, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. arXiv preprint arXiv: 1802.02611.

Cong L, Ding S, Wang L, Zhang A, Jia W (2018) Image segmentation algorithm based on superpixel clustering. IET Image Processing 12(11): 2030-2035.

Dai Y, Gao Y, Liu F (2021) Transmed: transformers advance multi-modal medical image classification. Diagnostics 11(8): 1384.

Guo P, Li N (2015) Automatic segmentation of cucumber disease images based on fuzzy clustering. Chinese Journal of Agricultural Chemistry (03): 123-126+131. https://doi.org/10.13733/j.jcam.issn.2095-5553.2015.03.031

Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (p. 7132-7141).

Huang M, Xu G, Li J, Huang J (2021) A method for segmenting disease lesions of maize leaves in real time using attention YOLACT. Agriculture 11: 1216. https://doi.org/10.3390/agriculture11121216

Hughes D, Salathé M (2015) An open access repository of images on plant health to enable the development of mobile disease diagnostics. arXiv preprint arXiv:1511.08060.

Huo F, Sun X, Ren W, Yang D, Yu T (2019) Improved k-means algorithm for color image segmentation in Lab space. Journal of Jilin University (Information Science Edition) (02): 148-154. https://doi.org/10.19292/j.cnki.jdxxp.2019.02.006.

Hussain N, Khan MA, Tariq U, Kadry S, Yar MAE, Mostafa AM, Alnuaim AA, Ahmad S, (2022) Multiclass cucumber leaf diseases recognition using best feature selection. Computers, Materials and Continua 70(2): 3281-3294.

Khan SD, Basalamah S, Lbath A (2023) Weed–Crop segmentation in drone images with a novel encoder–decoder framework enhanced via attention modules. Remote Sensing 15(23): 5615. https://doi.org/10.3390/rs15235615

Laborde D, Martin W, Swinnen J, Vos R (2020) COVID-19 risks to global food security. Science 369: 500–502. https://doi.org/10.1126/science.abc476

Liu L, Cheng X, Lai J (2018) A method for canopy image segmentation of cotton fields based on improved full convolutional networks. Journal of Agricultural Engineering (12): 193-201.

Liu Y, Hu L, Cao Y, Tang J, Lei B (2021a) U-Net-based algorithm for segmentation of maize leaf spots. Chinese Agronomy Bulletin (05): 88-95.

Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021b) Swin transformer: hierarchical vision transformer using shifted windows. arXiv preprint arXiv: 2103.14030.

Mi Z, Zhang X, Su J (2020) Wheat stripe rust grading by deep learning with attention mechanism and images from mobile devices. Frontiers in Plant Science 11: 558126.

Muhammed F (2022) Computer vision-based plant disease identification system: a review. Computer 1(1): 59-78.

Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. arXiv preprint arXiv: 1505.04597.

Russell BC, Torralba A, Murphy KP, Freeman WT (2008) LabelMe: a database and web-based tool for image. International Journal of Computer Vision 77(1).

Shelhamer E, Long J, Darrell T (2017) Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis & Machine Intelligence 39(04): 640-651.

Shi J, Malik J (2000) Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence 22(8): 888-905.

Tan P, Xiang H (2022) Research on transformer-based image segmentation. China New Technology and New Products (08): 23-26. https://doi.org/10.13612/j.cnki.cntp.2022.08.043

Ulutan O, Iftekhar ASM, Manjunath BS (2020) Vsgnet: spatial attention network for detecting human object interactions using graph convolutions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (p. 13617-13626).

Wang X, Guo X (2018) A method of green crop segmentation in agricultural fields based on GR color features. Heilongjiang Science (16): 14-15+19.

Wang X, Wang X, Liu Y, Guo X (2020) Segmentation of corn leaf diseases based on deep learning. Heilongjiang Science (20): 10-13.

Wang X, Yin L, Guo X (2018) Image segmentation method for green crops in agricultural fields under outdoor variable lighting conditions. Journal of Jilin University (Science Edition) (05): 1213-1218. https://doi.org/10.13413/j.cnki.jdxblxb.2018.05.28

Wang Z, Shi Y, Li Y (2019) Corn leaf disease spot segmentation based on improved full convolutional neural network. Computer Engineering and Applications (22): 127-132.

Wiesner-Hanks T, Wu H, Stewart E, DeChant C, Kaczmar N, Lipson H, Gore MA, Nelson RJ (2019) Millimeter-level plant disease detection from aerial photographs via deep learning and crowdsourced data. Frontiers in Plant Science 10: 1550.

Woo S, Park J, Lee JY, Kweon, IS (2018) Cbam: convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV) (p. 3-19).

Wu J, Yu R, Wang H (2021) A large-scale genomic association analysis identifies the candidate causal genes conferring stripe rust resistance under multiple field environments. Plant Biotechnology Journal 19(1): 177-191.

Xu W, Liu Y, Zhang H (2017) Research progress of image segmentation based on region growth. Beijing Biomedical Engineering (03): 317-322.

Yang L, Tian S, He X, Wang T, Wang B, Patel P, Yang X, et al. (2019) Ultrasound prostate segmentation based on multidirectional deeply supervised V-Net. Medical physics 46(7): 3194-3206.

Yuan H, Zhu J, Wang Q, Cheng M, Cai Z (2022) An improved DeepLab v3+ deep learning network applied to the segmentation of grape leaf black rot spots. Frontiers in Plant Science 13.

Zeng Q, Wu J, Liu S (2019) Genome-wide mapping for stripe rust resistance loci in common wheat cultivar qinnong 142. Plant Disease 103(3): 439-447.

Zeng Z (2022) Analysis of the current situation of corn production in China and suggestions. Grain, Oil and Feed Science and Technology (03): 4-8.

Zhang X, Xia Y, Xia N, Zhao Y (2020) Cotton image segmentation based on K-mean clustering with marker watersheds. Sensors and Microsystems (03): 147-149. https://doi.org/10.13873/J.1000-9787(2020)03-0147-03