

## Article

# EFS-Former: An Efficient Network for Fruit Tree Leaf Disease Segmentation and Severity Assessment

Donghui Jiang, Miao Sun, Shulong Li, Zhicheng Yang and Liying Cao \*

College of Information and Technology, Jilin Agricultural University, Changchun 130118, China; donghuijiang728@163.com (D.J.); sunmiao@mails.jlau.edu.cn (M.S.); lishulong@mails.jlau.edu.cn (S.L.); y1220223056@163.com (Z.Y.)

\* Correspondence: z2112020114@mails.jlau.edu.cn

**Abstract:** Fruit is a major source of vitamins, minerals, and dietary fiber in people's daily lives. Leaf diseases caused by climate change and other factors have significantly reduced fruit production. Deep learning methods for segmenting leaf diseases can effectively mitigate this issue. However, challenges such as leaf folding, jaggedness, and light shading make edge feature extraction difficult, affecting segmentation accuracy. To address these problems, this paper proposes a method based on EFS-Former. The expanded local detail (ELD) module extends the model's receptive field by expanding the convolution, better handling fine spots and effectively reducing information loss. H-attention reduces computational redundancy by superimposing multi-layer convolutions, significantly improving feature filtering. The parallel fusion architecture effectively utilizes the different feature extraction intervals of the convolutional neural network (CNN) and Transformer encoders, achieving comprehensive feature extraction and effectively fusing detailed and semantic information in the channel and spatial dimensions within the feature fusion module (FFM). Experiments show that, compared to DeepLabV3+, this method achieves 10.78%, 9.51%, 0.72%, and 8.00% higher scores for mean intersection over union (mIoU), mean pixel accuracy (mPA), accuracy (Acc), and F\_score, respectively, while having 1.78 M fewer total parameters and 0.32 G lower floating point operations per second (FLOPS). Additionally, it effectively calculates the ratio of leaf area occupied by spots. This method is also effective in calculating the disease period by analyzing the ratio of leaf area occupied by diseased spots. The method's overall performance is evaluated using mIoU, mPA, Acc, and F\_score metrics, achieving 88.60%, 93.49%, 98.60%, and 95.90%, respectively. In summary, this study offers an efficient and accurate method for fruit tree leaf spot segmentation, providing a solid foundation for the precise analysis of fruit tree leaves and spots, and supporting smart agriculture for precision pesticide spraying.

**Keywords:** H-attention; parallel fusion architecture; leaf diseases detection; smart agriculture; CNN



**Citation:** Jiang, D.; Sun, M.; Li, S.; Yang, Z.; Cao, L. EFS-Former: An Efficient Network for Fruit Tree Leaf Disease Segmentation and Severity Assessment. *Agronomy* **2024**, *14*, 1992. <https://doi.org/10.3390/agronomy14091992>

Academic Editor: Baohua Zhang

Received: 29 July 2024

Revised: 22 August 2024

Accepted: 30 August 2024

Published: 2 September 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

With the global increase in agricultural productivity, fruit farming has become a significant component of modern agriculture. However, changes in climatic factors and frequent disease outbreaks can damage photosynthesis through leaf diseases, weakening the vitality of fruit trees, which affects yield and threatens the economic interests of fruit growers and food safety for consumers [1]. The development of artificial intelligence has introduced new methods for precise pesticide application, significantly reducing pesticide use and economic costs for fruit farmers, while also decreasing chemical pollution of the land [2]. Thus, accurate crop disease analysis is essential for timely prevention. Traditional methods for detecting fruit diseases primarily rely on manual visual observation and judgment based on extensive experience. In large orchards with diverse fruits, this method is time-consuming, labor-intensive, and limited in accuracy by the inspector's expertise [3]. Advances in artificial intelligence have led to deep learning-based image processing technology, offering new solutions for fruit disease detection.

Computer image segmentation algorithms are widely used for analyzing leaves and diseases in agronomy. Researchers frequently use clustering, threshold segmentation, and region growing methods for leaf and disease segmentation. For instance, Febriyanto et al. [4] utilized the k-means clustering method for citrus leaf segmentation and disease detection, providing initial clusters specific to citrus. Chodey et al. [5] applied the fuzzy C-means method to segment target foreground and background, thereby identifying agricultural pests. Chodey et al. [6] detected cluttered backgrounds using advanced integrated color feature detection, which includes various color spaces, color indexes, and color to grayscale dialogs with singular value decomposition (SVD), and eliminated cluttered backgrounds by interactively selecting growing seeds in the advanced comprehensive color feature (ACCF) atlas using the region growing method. Ma et al. [7] used the interactive region growing method in the comprehensive color feature (CCF) atlas to segment vegetable foliar lesion images taken in greenhouses with complex backgrounds. Ma et al. [8] applied non-local mean filtering and two-dimensional histograms to denoise maize disease images, and improved the integrated particle swarm optimization (PSO) method with a new elite strategy for precise segmentation of various maize leaf diseases. Z. Xie et al. [9] integrated the Otsu method with the innovative dung beetle optimizer (DBO) to achieve excellent performance in real and complex rubber image segmentation tasks. These methods are often specific to certain diseases and rely on large amounts of high-quality data; they may not perform well if there are insufficient or non-representative training data. Typically, these methods perform poorly when applied to different crop diseases, limiting their broader application in agriculture.

With ongoing advancements in deep learning technology, semantic segmentation has become a crucial application. Fully convolutional network (FCN) was the first model to utilize a fully convolutional network for semantic segmentation [10]. It achieves image-to-image prediction by replacing the fully connected layer with a convolutional layer and introducing skip connections to fuse different levels of information, enabling pixel-level classification and establishing a solid foundation for semantic segmentation technology. U-Net, originally designed for biomedical image segmentation, has a symmetric encoder-decoder structure resembling the letter U. Its features are combined with decoder features through skip connections. It performs well on small sample data by combining encoder and decoder features through skip connections [11]. PSPNet fuses feature maps at different scales using the pyramid pooling module to extract multi-scale contextual information, enabling segmentation of targets at various scales [12]. Other CNN-based semantic segmentation networks include DeepLab [13], SegNet [14], and Mask R-CNN [15]. Due to their strong transferability and high accuracy, these models have been increasingly introduced into agriculture, yielding good results. Jia et al. [16] improved Mask R-CNN to effectively address fruit overlapping in real environments. ResNet and DenseNet reduce the number of parameters and serve as backbone networks for feature extraction, then are embedded into edge devices for field testing, supporting the development of picking robots. Zou et al. [17] proposed a novel image enhancement method and a simplified U-Net network for weed segmentation in complex field scenes. This method excelled in single image processing speed and had an average intersection over union (IoU) ratio. Kang et al. [18] proposed the method's effectiveness in cotton root segmentation, and demonstrated this by incorporating the attention mechanism into the DeepLabv3+ model and comparing it with SegNet and U-Net. Azizi et al. [19] used the VGG16 deep model for plot segmentation alongside the watershed segmentation method. C. Wang et al. [20] divided the cucumber disease segmentation task into two parts: segmenting cucumber leaves using DeepLabv3+ and extracting disease spots using U-Net. This method simplifies the segmentation task and achieves higher point segmentation accuracy. Sunil et al. [21] extracted the features of healthy and diseased leaves of tomatoes through multilevel attention from various aspects such as spatial, channel, and pixel, which in turn gave good results. B.-Y. Liu et al. [22] used two-level CNNs, PSPNet and U-Net, to accurately segment apple leaves and spots and accurately derive the severity of the disease through the area ratio

between pixels. Dai et al. [23] used meteorological data enhancement methods to effectively simulate the climatic conditions of a real orchard, while the use of multilevel attention and global average pooling (GAP) layers reduced computational costs and achieved good segmentation accuracy. Despite the success of previous research, limitations remain. Due to receptive fields and shared parameters, CNNs excel at local feature extraction through convolution and pooling operations but lack the ability to capture global information and long-range dependencies. Some studies propose effective solutions, such as expanding the receptive field or adopting deeper network architectures. However, deeper layers exacerbate network degradation [24]. Thus, CNN-based segmentation networks struggle to balance computational complexity and segmentation accuracy.

In recent years, the Transformer [25], based on the self-attention mechanism and fully connected layers, has been introduced to computer vision, showing strong generalization due to its global nature and ability to capture long-range dependencies. Some researchers initially combined the attention mechanism with traditional convolution. For instance, X. Wang et al. [26] proposed non-local neural networks (NLNs), which use non-local operations to capture long-range dependencies. H. Zhang et al. [27] introduced a self-attention mechanism in generative adversarial networks, significantly enhancing the quality of generated images. This approach combines the attention mechanism with convolutional layers. Dosovitskiy et al. [28] first used the self-attention mechanism in place of traditional convolution, significantly enhancing the performance of a pure attention model through large-scale pre-training and slice embedding methods, making it feasible and efficient for practical applications. Others have developed corresponding segmentation models, such as Segmentation Transformer (SETR) [29], SegFormer [30], and Swin Transformer [31]. Compared to CNN, Transformers divide images into equal-sized patches and use a self-attention mechanism to capture long-range dependencies and spatial transformation relationships that reflect global properties, enabling global modeling. Although Transformers perform well in the above areas, they have shortcomings. For example, the self-attention mechanism in the encoder redundantly calculates neighboring pixel correlations and tends to be inferior to CNN in extracting local details. Consequently, the decoder's inability to recover original information during up-sampling can degrade performance in tasks requiring fine segmentation.

Taking into account the overall generalization of the model and the accuracy of segmentation in real-world environments (such as difficulty in extracting edges due to the influence of light on fruit tree leaves, and the easy omission of local features due to the small target size of disease points), this article proposes a complementary network that integrates the Transformer and CNN, where CNN complements local feature extraction and Transformer globally, and designs EFS-Former for the segmentation of various fruit leaf diseases. Among them, the Transformer block performs position encoding and global modeling to efficiently extract global and positional features of leaves and diseases. The CNN module optimizes the local details and edge information, which improves the pixel classification ability of the model, resulting in finer segmentation results of leaves and spots, and extracts more information of tiny spots. In addition, we introduce a feature fusion module that fuses the obtained feature information from both branches. A richer feature representation is obtained through feature merging and subsequent refinement via the remaining layers.





## 2. Materials and Methods

### 2.1. Dataset

In this work, four diseases of three fruit crops were collected, namely, apple spotted leaf drop disease, pomegranate cercospora spot, grape brown spot, and black rot. We classified grape brown spot and black rot into early and late stages of the disease using the size of the leaf area occupied by disease spots; apple spotted leaf drop disease was classified into indoor and outdoor scenarios for segmentation; images taken with different light intensities and backgrounds were also collected to enrich the outdoor dataset. All the

above datasets were obtained from Plant Village [32], a publicly available dataset for crop pest and disease identification. They were collected and labeled by us. As the pixel size was not uniform across the images when they were collected, the image size was subsequently adjusted to  $512 \times 512$  pixels, and the types of images included in this dataset are listed in Table 1.

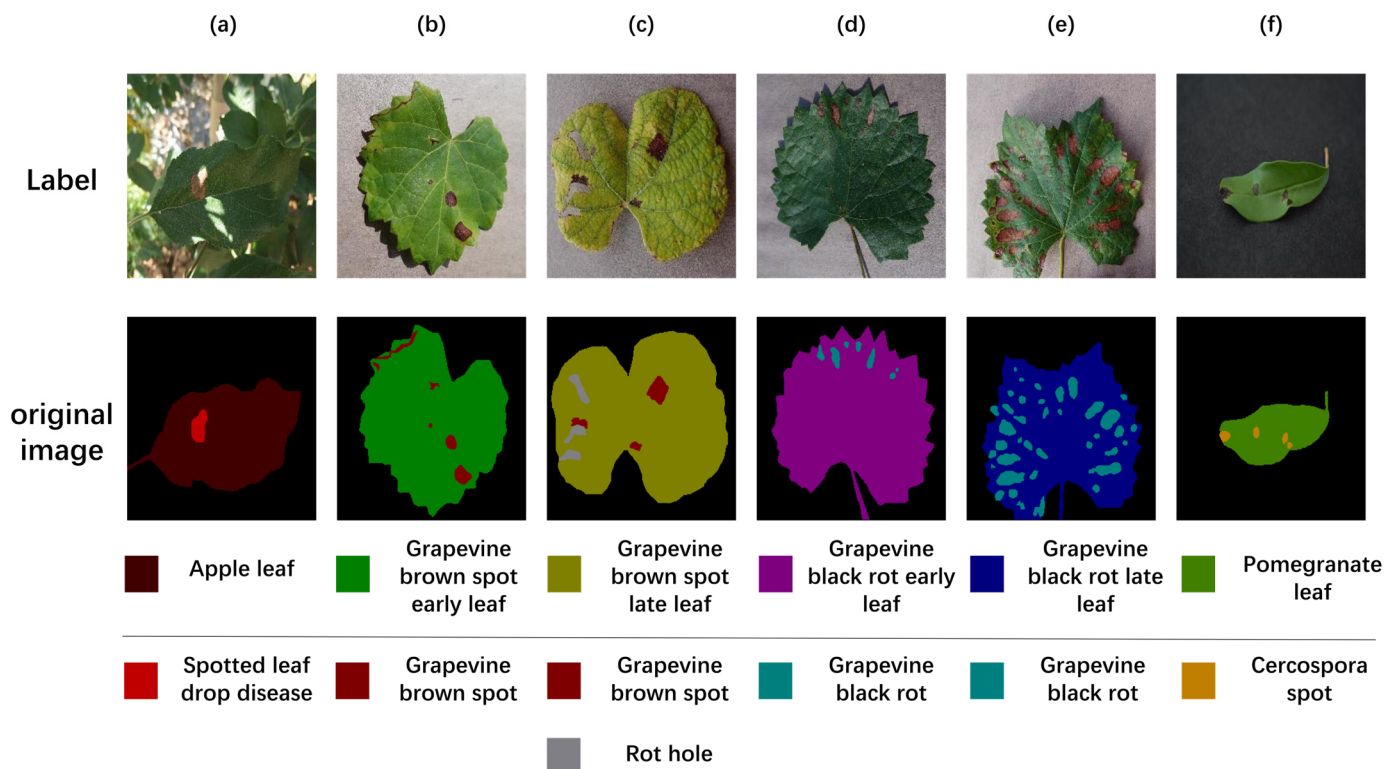
**Table 1.** Samples of four types of fruit and leaf diseases and hazards.

Fruit and Leaf Type	Type of Disease	Original Image	Hazards	Challenges
Grape	Brown spot		Disrupts leaf photosynthesis and triggers early defoliation.	The image is disturbed by background noise (background blur), which blurs the blade edge information, and can affect the segmentation of the model.
Grape	Black rot		It will increase the population of pathogenic bacteria.	Leaf blade edges are curled, resulting in poorly defined edge features.
Pomegranate	Cercospora spot		It increases the rate of diseased leaves, and early leaf fall is obvious, which is unfavorable to flower bud differentiation.	Darker lighting conditions result in spot features that are similar to the background information, making feature extraction difficult.
Apple	Spotted leaf drop disease		Causes early defoliation, weakens fruit trees and affects fruit production.	Due to different lighting conditions during image acquisition, the reflected light on the surface of leaves or fruits may cause uneven brightness, increasing the difficulty of segmentation.

As shown in Table 1, the data demonstrate the challenges when segmenting different pathologies of various fruits. For fruits and leaves, the challenges are as follows: (1) poorly defined leaf edges due to curling; (2) light affecting fruit and leaf edges, making edge feature extraction difficult; and (3) multiple overlapping leaves in outdoor scenes create a complex background, complicating global feature extraction and greatly affecting model segmentation accuracy. For disease spots, the challenges are as follows: (1) targets are too small and easily missed during local feature extraction; (2) diseased points are too illuminated, complicating feature extraction; and (3) blurred contours of disease points or pixel points close to leaves hinder model feature extraction.

## 2.2. Image Preprocessing

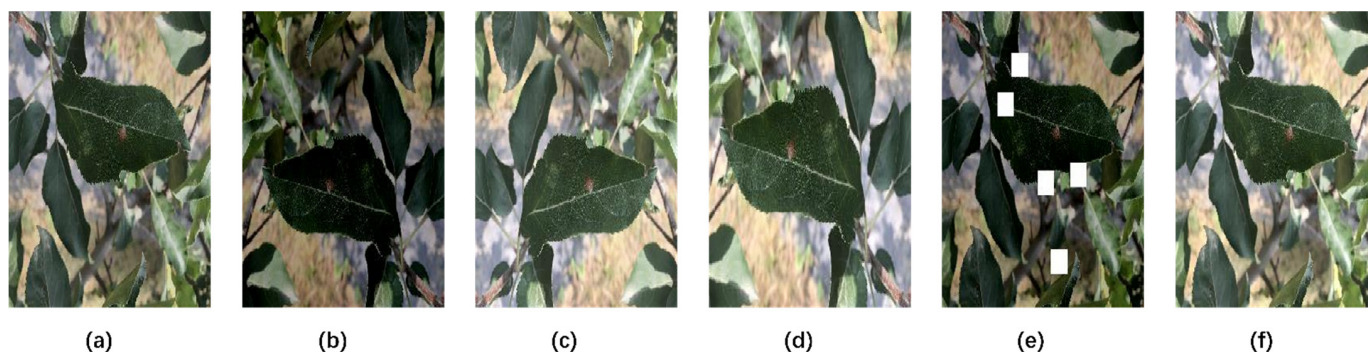
This work sums up the experience of previous work, using the Labelme annotation software [33]; the original dataset annotation work was used here for model learning and the evaluation of the model of the standard. The entire annotation dataset includes three fruit crops of four diseases, totaling twelve categories. For grape black rot and grape brown spot disease, the early and late disease stages are classified using the size of the disease spots area. In the annotations of the apple spotted leaf drop disease outdoor images, not only single leaves but diseased leaves are divided into separate annotation categories to ensure that the model can be better adapted to the complex outdoor environment. The original images of the dataset and the labeled images are displayed in Figure 1. Among them, (a) shows apple spotted leaf drop disease, (b) and (c) show the early and late stages of grape brown spot disease, (d) and (e) show the early and late stages of grape black rot disease, and (f) shows pomegranate cercospora spot, respectively. The categories represented by each color are shown separately in the figure.



**Figure 1.** Original and annotated images: (a) apple spotted leaf drop disease, (b) early stage of grape brown spot disease, (c) late stage of grape brown spot disease, (d) early stage of grape black rot disease, (e) late stage of grape black rot disease, and (f) pomegranate cercospora spot.

Neural networks require a large amount of sample data for training. A small sample size can lead to underfitting, preventing the network from completing training. Therefore,

expanding the original dataset is essential. In this study, the dataset was expanded by a ratio of 1:5. This included randomly flipping and zooming images, adjusting brightness and contrast, and adding  $20 \times 20$  white mask blocks. These enhancement methods simulated issues such as varying light conditions and leaf overlapping that may occur during data collection. Taking apple spotted leaf drop disease as an example, the enhanced images are shown in Figure 2. The specific numbers of different categories of leaves are listed in Table 2.



**Figure 2.** Original image and enhanced image: (a) original image, (b) reduced brightness and flipped, (c) randomly flipped, (d) random zoom, (e) white mask block added, (f) panning and increasing brightness.

**Table 2.** Number of pictures per fruit condition.

	Indoor Pictures			Outdoor Pictures	
	Apple Spotted Leaf Drop Disease	Grape Black Rot	Pomegranate Cercospora Spot	Grape Brown Spot	Apple Spotted Leaf Drop Disease
Original	157	677	271	866	335
Enhanced	785	3385	1355	4330	1675
Total	942	4062	1626	5196	2010

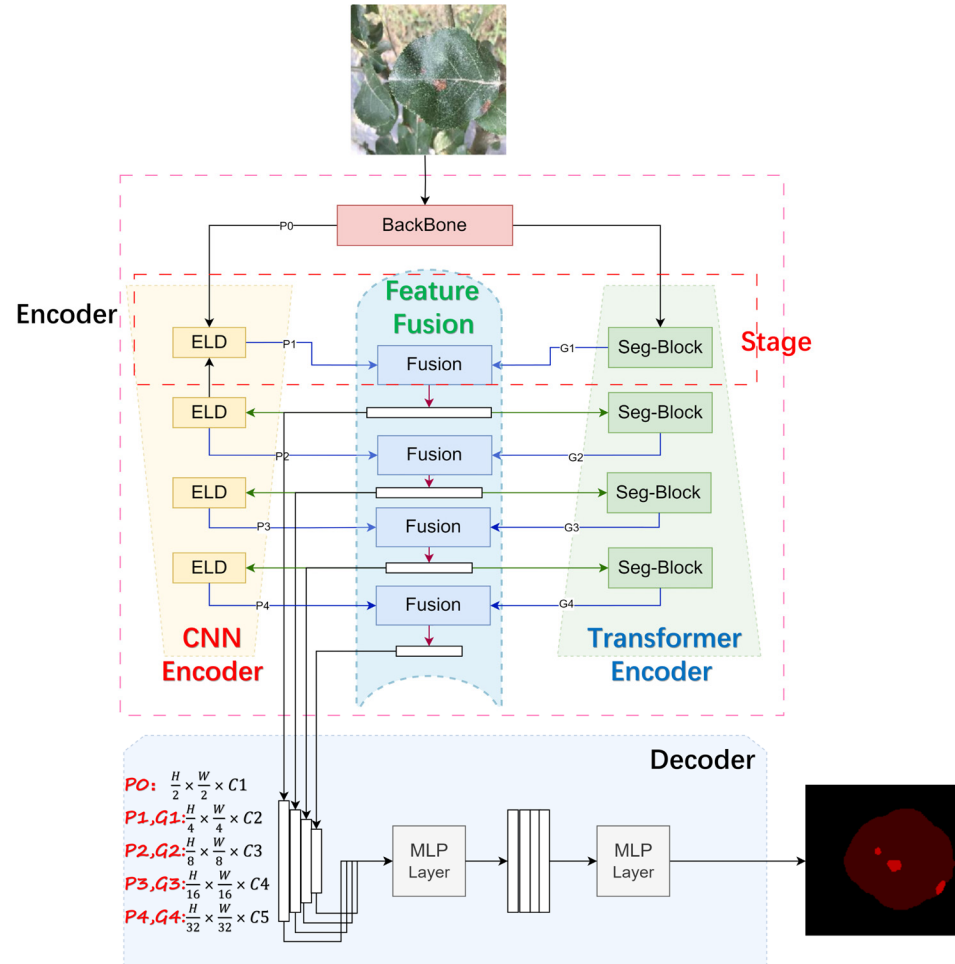
### 2.3. Methods

Inspired by the powerful representation capabilities of CNN and Transformer, we propose the CNN and Transformer Complementary Network (EFS-Former). As shown in Figure 3, the main difference between our approach and existing architectures is our design of two different encoders to generate complementary features, along with cross-domain complementary fusion and multi-scale feature fusion.

Our motivation for using both CNN and Transformer encoders is that they complement each other. CNN excels at local feature extraction, while Transformer efficiently captures long-range dependencies. Due to the small traits of agricultural diseases, deeper CNN architectures can lose these smaller traits. It is difficult for the decoder to recover these lost small and narrow targets, resulting in failure. Conversely, the Transformer encoder can capture more dependencies and efficiently recover many features in the decoder section.

Specifically, the input image for the network is sized  $512 \times 512 \times 3$ . First, the image is input to the backbone network for initial feature extraction, producing features mapped to the original image size  $\frac{1}{2}$ , denoted as P0. This step effectively reduces the number of initial parameters for the encoder stage. The entire encoder is divided into four stages, each stacked four times using the same parallel fusion structure. This deepens the network layers to continuously refine fine features, minimizing missed features. The feature map from the backbone network is input to both CNN and Transformer encoders for extraction in different regions. The Transformer stage captures fine features in the global context, generating feature images G1, G2, G3, and G4. The CNN stage mainly captures local fine features, obtaining multi-level feature maps P1, P2, P3, and P4. The feature maps generated

by both encoders are of the original image’s  $\{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$  resolution. Each stage outputs  $\{16, 32, 64, 160, 256\}$  channel dimensions  $C_i$  ( $i = 1, 2, 3, 4, 5$ ). In the decoder section, the feature maps P1–P4 from the encoder, downsampled at a  $2\times$  rate, are fused and upsampled to the original size, producing  $512 \times 512 \times Ncls$  segmentation results.  $Ncls$  is the number of categories, which is 3 in this paper.

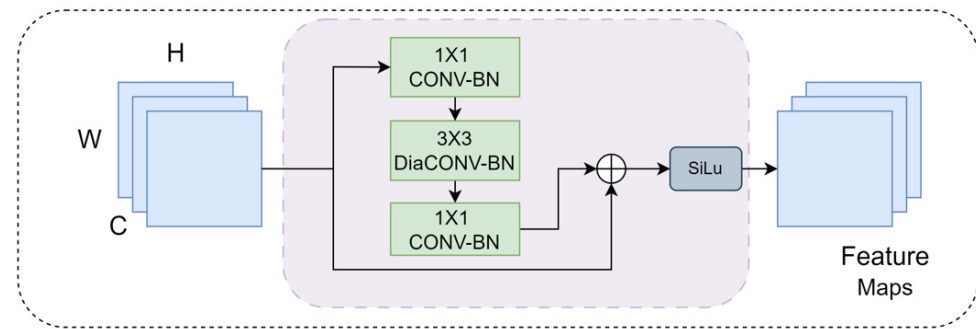


**Figure 3.** The overall architecture of EFS-Former, including the main structure of parallel fusion, CNN encoder, feature fusion module, and improved Transformer encoder.

### 2.3.1. Expanded Local Detail (ELD) Module

The curled and jagged shape of leaf edges and shadows from different lighting angles reduce leaf segmentation accuracy. Additionally, the similarity between the edges of diseased points and leaf color makes extracting diseased points more difficult, leading to loss of details and reduced segmentation accuracy. Furthermore, in most fruit and leaf images, diseased pixels comprise a small proportion of the total image, making it harder to extract features with small diseased spots. Therefore, this subsection leverages CNN’s local feature extraction and designs an expanded local detail (ELD) module to optimize the edge segmentation of leaves and spots and to extract more subtle spots. The ELD module is shown in Figure 4.

While the Transformer module captures remote feature dependencies, it often loses the relationship between local features. Conversely, CNN clearly obtains local feature relationships through the translation of the convolutional kernel with a specific step size, enabling feature extraction with a stronger dependency on texture features and effectively protecting local feature extraction.



**Figure 4.** ELD module architecture diagram.

As shown in Figure 4, the ELD module first uses  $1 \times 1$  convolution with sigmoid linear unit (SiLU) activation function to introduce non-linear transformations, change the channel dimensions, and perform dimensionality upgrading operations to make the network learn deeper features, and then uses the  $3 \times 3$  convolution with an expansion rate of 2 to reduce the computational volume of the network without increasing the number of convolution kernels to expand the feeling of the module and capture the inputs better in the global context. It then uses the  $1 \times 1$  convolution to reorganize and integrate the channel information of the input feature map. Finally, the  $1 \times 1$  convolution is used to reorganize and integrate the channel information of the input feature map, and, through the batch normalization and SiLU activation functions, the network can introduce non-linear transformations to improve the expression and differentiation of the features in the model, and then finally downscale to the original dimensions to better adapt to the fusion operation in the next stage. Assuming  $x$  as the input, the features that go through the ELD layer ( $y_i$ ) each time can be expressed as follows:

$$y_{out} = SiLU(concat(x, BN_3(W_3 * (BN_2(W_2 *_d (BN_1(W_1 * x))))))) \quad (1)$$

where  $y_{out}$  denotes the feature output after the ELD layer,  $*$  denotes the standard convolution operation,  $*_d$  denotes the dilation convolution operation,  $BN_i(z)$  denotes the BatchNorm operation at layer  $i$ ,  $concat(x, y_3)$  denotes the splicing operation in the channel dimension, and SiLU denotes the SiLU activation function.

Therefore, the ELD module effectively enhances the perception of leaf and spot edge features by locally optimizing patches embedded in the output feature layer. It interacts with the global Seg-Blok module to perform coarse and fine feature fusion in the feature fusion module, thus improving the extraction of tiny diseases and segmentation of leaf and spot edges.

### 2.3.2. Seg-Block

In image segmentation tasks, the receptive field is crucial as it reflects the perceptual range of the convolutional kernel on the feature map. A small receptive field can only capture unilateral local information. In agricultural disease segmentation, where diseases are small and leaf edges can be blurred, researchers focus on these challenges. Our network enhances the encoder with a parallel Transformer and an attention mechanism using selective convolution, aiming to increase the receptive field while preserving detailed features. The Transformer uses an attention mechanism to perform a dot product (or other similarity measure) between the query and key vectors to calculate attention scores, which represent the correlation between the query and key, thereby enabling global modeling. However, the attention mechanism is computationally intensive and inevitably leads to computational redundancy. Additionally, due to close attention scores, certain features may be lost, negatively impacting segmentation results. To address these issues, this paper proposes a Transformer module (Seg-Block) that integrates H-attention, a feedforward neural network (FNN), and a patch merging module. This module aims to enhance the



performance and computational efficiency of traditional Transformer models by combining global and local feature extraction capabilities.

As shown in Figure 5, H-attention helps the model to be more flexible and accurate in processing information with different spatial resolutions by incorporating convolutional layers with different sizes of convolutional kernels in the multi-attention module. To further enhance the feature representation capability, the input features are reshaped, and then multiple convolutional layers with different kernel sizes are passed through for multi-scale feature extraction. These convolutional layers use different convolutional kernel sizes (e.g.,  $1 \times 1, 3 \times 3, 5 \times 5, 7 \times 7$ ) to capture spatial information at different scales.

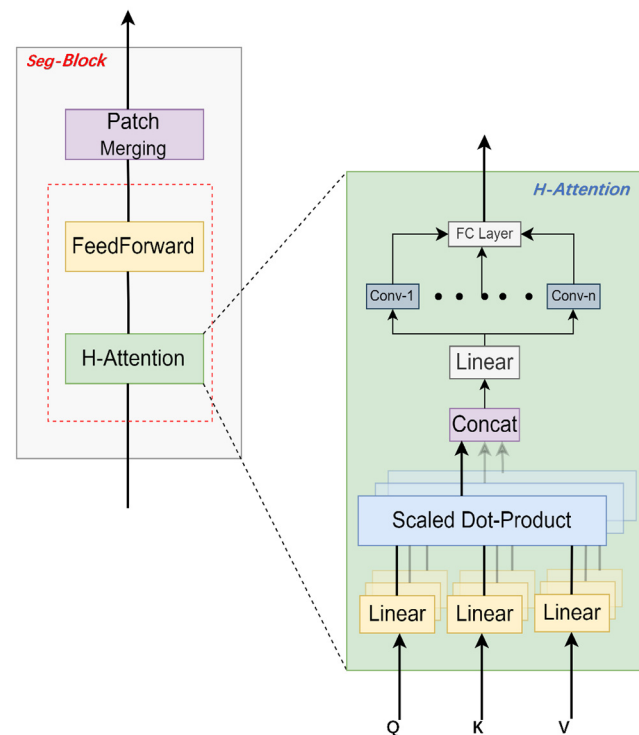


Figure 5. Seg-Block architecture diagram.

The output features of each convolutional layer are accumulated to obtain an overall feature map, which is compressed through a fully connected layer to generate a feature vector that is used to compute selective weights. The feature vector is reprojected back to the original dimension through a series of fully connected layers, and these projections are used to generate selective attentional weights for multiscale features. The weights are normalized by a softmax function and multiplied by the corresponding convolutional output features, and, finally, the features at different scales are weighted and summed to generate an enhanced feature map. After a series of fine-grained operations on the H-attention input features, the output feature map retains more useful information, while improving the processing capability for complex visual tasks. The overall operation flow is as follows:

The raw image input first-stage attention mechanism is as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

where  $Q$  is the query matrix,  $K$  is the key matrix,  $V$  is the value matrix, and  $d_k$  is the dimension of each header.

The second stage of the operation is the input features after multiple convolutions and operations; the convolution output features are obtained and the convolution stack is summed. Subsequently, the channel attention weights are calculated and, finally, the

weights are multiplied with the convolution output in a multiplication operation. The overall computational procedure for the second stage is as follows:

$$X_{conv}^{(i)} = Conv^{(i)}(X), i = 1, \dots, k \quad (3)$$

$$U = \sum_{i=1}^k X_{conv}^{(i)} \quad (4)$$

$$S = \frac{1}{HW} \sum_{h=1}^H \sum_{w=1}^W U_{hw} \quad (5)$$

$$Z = ReLU(W_1 S) \quad (6)$$

$$w_{conv}^{(i)} = Softmax(W_2^{(i)} Z), i = 1, \dots, k \quad (7)$$

$$V = \sum_{i=1}^k w_{conv}^{(i)} \odot X_{conv}^{(i)} \quad (8)$$

$$X_{final} = Attention(Q, K, V) \quad (9)$$

$$X_{final} = H - Attention(Q, K, V) + V \quad (10)$$

where, after the convolutional output of the  $i$ th convolutional kernel,  $k$  is the number of convolutional kernels,  $U$  is the summation of all the convolutional outputs,  $H$  is the feature height,  $W$  is the feature width,  $S$  represents the global average pooling result of the feature map after the selective convolutional attention module,  $Z$  is the intermediate feature after activation of rectified linear unit(ReLU),  $W_1$  is the weight of the first fully connected layer,  $w_{conv}^{(i)}$  is the  $i$ th convolutional kernel channel attentional weights, and  $W_2^{(i)}$  is the  $i$ th convolution kernel second fully connected layer weights.

The FNN first processes the input features using a linear transformation (fully connected layer). The linear transformation maps them to a higher dimensional space, typically extending the dimensionality by  $4 \times$ . This expansion operation allows the network to perform complex feature learning in higher dimensions. The features activated by GELU are then subjected to a second linear transformation that reduces the feature dimensions to the original ones. This reduction operation ensures that the output of the FNN has the same dimensions as the input, allowing it to be seamlessly integrated with subsequent operational modules. Residual connection and layer normalization are also introduced in the FNN. Residual connection allows the input features to be directly added to the output of the FNN, thus enhancing gradient flow. Layer normalization provides better stability of the model across training batches by normalizing each layer of the input features.

Patch merging progressively reduces the spatial resolution of the feature map by merging neighboring image chunks, merging adjacent  $2 \times 2$  chunks, and mapping the merged features to a new feature space through a linear projection layer. By merging neighboring chunks, the model can aggregate more feature information while reducing computational effort. This is important for capturing multi-scale contextual information and reducing computational complexity.

### 2.3.3. Feature Fusion Module

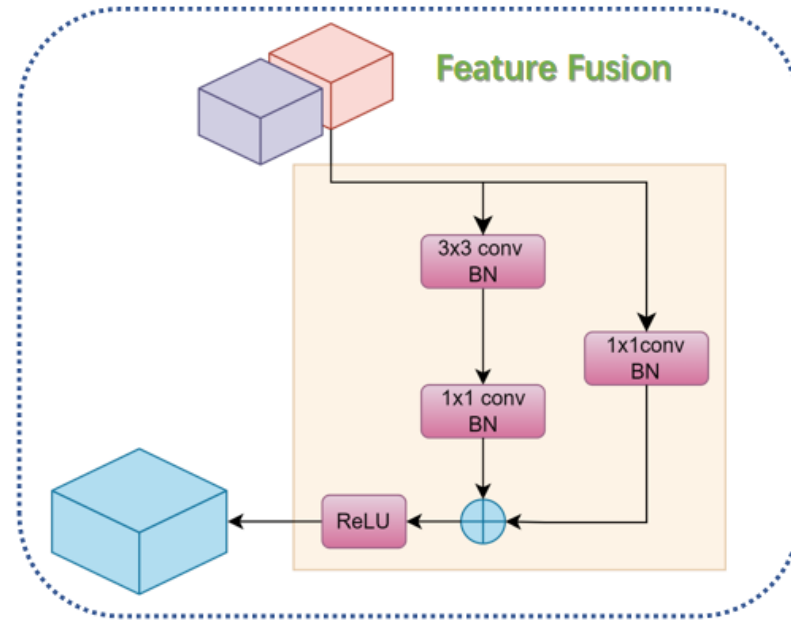
Since Transformer and CNN encoders are designed for different tasks, they employ distinct feature extraction methods and have different application purposes. Transformer and CNN encoders have different intervals of interest or receptive fields for the same input. In order to obtain complementary information, we design a lightweight feature fusion module (FFM), as shown in Figure 6 and the following calculation formula. The feature maps output from the Transformer and CNN encoders are first spliced in the channel dimension to obtain a combined feature map. This feature map contains complementary information from the two encoders and provides the basis for subsequent fusion operations, as follows:

$$X_T \in R^{B \times C_T \times H \times W} \quad (11)$$

$$X_C \in R^{B \times C_C \times H \times W} \quad (12)$$

$$X_{concat} = Concat(X_T, X_C) \in R^{B \times (C_T + C_C) \times H \times W} \quad (13)$$

where  $X_T$  represents the feature map obtained from the Transformer encoder, and  $B$  is the batch size.  $C_T$  represents the number of channels output by the Transformer encoder.  $W$  is the height and width of the feature map. The feature map  $C_C$  obtained from the CNN encoder is the number of channels output by the encoder.  $X_{concat}$  is the concatenated feature map.



**Figure 6.** Overall architecture of FFM.

Subsequently, spatial fusion through  $3 \times 3$  convolution can extract more complex and diverse local features while preserving the information of the original image. The  $3 \times 3$  convolution extracts local information in the spatial dimension of the feature map through a sliding window operation, while preserving the spatial structure of the original features. The output feature map  $X_{3 \times 3}$  has  $C_{out}$  channels, reflecting the fusion and extraction of spatial features, as follows:

$$X_{3 \times 3} = BN(Conv_{3 \times 3}(X_{concat})) \in R^{B \times C_{out} \times H \times W} \quad (14)$$

where  $X_{3 \times 3}$  refers to the feature map after  $3 \times 3$  convolution.  $C_{out}$  is the number of channels in the  $3 \times 3$  convolution output, which is usually an adjustable parameter to control the dimension of the output features.

The final  $1 \times 1$  convolution acts on the channel dimension of the feature map, which corresponds to a linear combination of channels for each pixel point. This operation enables features between different channels to be recombined and weighted, resulting in a richer representation of channel features. The combination of batch normalization and ReLU activation function enables the  $1 \times 1$  convolution to perform a non-linear transformation of the features, which improves the expressive power of the model, as follows:

$$X_{1 \times 1} = BN(Conv_{1 \times 1}(X_{3 \times 3})) \in R^{B \times C_{out} \times H \times W} \quad (15)$$

$$X_{res1 \times 1} = BN(Conv_{1 \times 1}(X_{concat})) \in R^{B \times (C_T + C_C) \times H \times W} \quad (16)$$

$$X_{concat1} = Concat(X_{res1 \times 1}, X_{1 \times 1}) \in R^{B \times C_{out} \times H \times W} \quad (17)$$

$$X_{out} = ReLu(X_{concat1}) \quad (18)$$

$X_{1 \times 1}$  is the feature map after a  $1 \times 1$  convolution,  $X_{res1 \times 1}$  refers to the feature map connected by residuals, and  $X_{concat1}$  represents the feature map after the second concatenation.

#### 2.4. Experimental Platform and Experimental Evaluation Indices

To verify the performance of the EFS-Former and its various modules proposed in this article, all experiments were conducted under the same software and hardware conditions, with the following operating conditions. The software platform was based on Python 3.8.19, with its main libraries including PyTorch 2.2.2 for model construction and training, OpenCV 4.1.2 for image processing and segmentation, and NumPy 1.24.4 for mathematical computation. These library versions were chosen for their compatibility and stability on the hardware platform. Experiments were conducted on Ubuntu 22.04.4 Long-Term Support (LTS) with an 11th Gen Intel® Core™ i5-11400F @ 2.60 GHz  $\times$  12 and an NVIDIA GeForce Ray Tracing Texel eXtreme (RTX) 2080 Ti graphics card. The optimal training hyperparameters were selected after repeated experiments: the optimizer was AdamW with weight decay, using a cosine learning rate decay strategy, with momentum of 0.9, a weight decay of  $10^{-2}$ , and a batch size of 6200 epochs. For faster convergence to a more refined convergence process, we set the initial learning rate to be  $10^{-4}$  and the minimum learning rate to be  $10^{-7}$ . This range allowed for larger weight updates in the initial phase so that the model quickly approached the optimal solution, and then progressively refined the updates to ensure final convergence. To prevent overfitting from occurring, we set the drop path rate to 0.1. For training weights, we used self-training weights for our experiments.

We used eight evaluation metrics to assess the proposed model's effectiveness: mean intersection over union (*mIoU*), mean pixel accuracy (*mPA*), accuracy (*Acc*), *F\_score*, *recall*, *precision*, *FLOPs*, and *total parameters*. These metrics evaluate the model from various aspects.

*mIoU*: In image segmentation tasks, the model assigns each pixel to a predefined category. *mIoU* measures the overlap between model predictions and true labels, calculated as follows:

$$IoU = \frac{TP_i}{TP_i + FP_i + FN_i} \quad (19)$$

$$mIoU = \frac{1}{N} \sum_{i=1}^N \frac{TP_i}{TP_i + FP_i + FN_i} \quad (20)$$

where  $N$  is the number of categories,  $TP$  is the true positive (pixels correctly predicted as positive),  $TN$  is the true negative (pixels correctly predicted as negative),  $FP$  is the false positive (pixels incorrectly predicted as positive), and  $FN$  is the false negative (pixels incorrectly predicted as negative).

*mPA* is calculated by making a binary classification judgment for each pixel and dividing the number of correctly predicted pixels by the total number of pixels. It measures the proportion of correctly classified pixels, as follows:

$$mPA = \frac{\sum_{i=1}^N TP_i}{\sum_{i=1}^N (TP_i + FP_i)} \quad (21)$$

*Accuracy* is the ratio of the number of samples correctly predicted by the model across all samples to the total number of samples. It measures the overall classification accuracy of the model, i.e., the number of samples correctly predicted by the model as a proportion of the total number of samples. The specific formula is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (22)$$

*F\_score* combines *precision* and *recall* to evaluate binary or multiclassification models. It is calculated as a weighted average of precision and recall. Including *F\_score* as a metric provides a comprehensive view of model performance, especially when precision and *recall* are equally important. The calculations are as follows:

$$Precision = \frac{TP}{TP + FP} \quad (23)$$

$$Recall = \frac{TP}{TP+FN} \quad (24)$$

$$F_{score} = (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \quad (25)$$

where  $\beta$  is a parameter used to regulate the relative importance of precision and recall. When  $\beta = 1$ , it is the common *F1-score*, which balances the weights of precision and recall.

### 3. Results and Discussion

In this study, we will validate the fruit tree leaf spot segmentation model based on the EFS-Former method, aiming to address the limitations of traditional deep learning methods when facing complex leaf features. The hypotheses are as follows: (1) The EFS-Former method can effectively improve the segmentation accuracy of the model for fruit tree leaf spots by introducing the ELD (expanded local detail) module and the H-attention mechanism. (2) The parallel fusion architecture can combine the advantages of the two encoders, CNN and Transformer, to achieve the accurate extraction of the edge features of fruit tree leaves, to reduce the impact of leaf folding, jaggedness, light shading, and other factors on the segmentation accuracy. (3) By calculating the proportion of leaf area occupied by diseased spots, this method can accurately determine the period of disease and provide key data support for fruit tree disease management.

#### 3.1. Analysis of Experimental Results

##### 3.1.1. Different Models

This subsection compares the models presented in this paper with several current state-of-the-art and classical semantic segmentation methods, including PSPNet, HRNetV2 [34], U-Net, DeepLabv3+ [35], SegFormer, and FCN. Our focus is on the clarity of the contour between the leaves and the target lesion points, as well as the segmentation of small target points. Each method was trained and tested on indoor and outdoor datasets containing four diseases of three fruits. Eight metrics, namely mIoU, mPA, recall, F\_score, Acc, precision, FLOPs, and total parameters, were used to measure the feasibility and effectiveness of our proposed models. The details are shown in Tables 3–5.

**Table 3.** Indicators for early and late assessment of grape brown spot under different models (in the table, (early stage) and (end stage) represent the period of leaf disease).

Method	Grape Brown Spot											
	mIoU			mPA			Precision			Recall		
	Early Stage	End Stage	Disease Spot	EARLY STAGE	End Stage	Disease Spot	Early Stage	End Stage	Disease Spot	Early Stage	End Stage	Disease Spot
PSPNet	91.0%	92.0%	40.0%	95.0%	97.0%	48.0%	95.0%	95.0%	95.0%	95.0%	97.0%	48.0%
HRNetV2	95.0%	94.0%	77.0%	97.0%	97.0%	85.0%	97.0%	97.0%	89.0%	97.0%	97.0%	85.0%
U-Net	90.0%	90.0%	77.0%	94.0%	96.0%	85.0%	95.0%	94.0%	89.0%	94.0%	96.0%	85.0%
DeepLabv3+	93.0%	92.0%	80.0%	95.0%	97.0%	88.0%	98.0%	95.0%	89.0%	95.0%	97.0%	88.0%
SegFormer	90.0%	91.0%	73.0%	94.0%	93.0%	79.0%	96.0%	97.0%	90.0%	95.0%	97.0%	80.0%
FCN	93.0%	93.0%	78.0%	98.0%	96.0%	50.0%	94.0%	90.0%	88.0%	95.0%	94.0%	55.0%
Ours	96.0%	96.0%	85.0%	98.0%	98.0%	89.0%	98.0%	98.0%	94.0%	98.0%	98.0%	89.0%

Table 3 shows that the method exhibits good segmentation performance for both early and late periods of grape brown spot. It divides the disease stages by the proportion of leaf area occupied by diseased spots and performs significantly better than other models in this aspect. Compared to SegFormer, the mIoU for early and late leaf segmentation and spot segmentation were 6%, 5%, and 7% higher, respectively. The mIoU for early and late leaf segmentation and spot segmentation by HRNetV2 were 1%, 2%, and 8% lower than the proposed method, respectively. The above results indicate that this method accurately divides the early and late stages of grape brown spot disease, and shows significant advantages in comparing different models. These results indicate that this method has great potential in practical applications and is of great significance for the precise spraying

of pesticides in orchards. It can help improve agricultural production efficiency and reduce pesticide use, providing strong support for scientific management during disease periods.

**Table 4.** Indicators for the assessment of pomegranate cercospora spot and apple spotted leaf drop disease under different models.

Method	Pomegranate Cercospora Spot					Apple Spotted Leaf Drop Disease				
	mIoU		mPA	Precision	Recall	mIoU		mPA	Precision	Recall
	Leaf	Disease				Leaf	Disease			
PSPNet	65.00%	29.00%	51.00%	80.00%	51.00%	83.00%	39.00%	62.50%	84.00%	64.50%
HRNetV2	96.00%	72.00%	89.50%	92.50%	89.50%	97.00%	72.00%	90.50%	91.00%	90.50%
U-Net	86.00%	67.00%	89.00%	89.00%	79.00%	82.00%	65.00%	85.00%	89.50%	85.00%
DeepLabv3+	96.00%	68.00%	90.50%	88.50%	90.50%	86.00%	65.00%	86.50%	85.50%	86.50%
SegFormer	95.00%	73.00%	90.00%	92.50%	90.00%	95.00%	69.00%	88.00%	91.00%	88.00%
FCN	95.00%	66.00%	82.00%	96.50%	82.00%	86.00%	70.00%	72.00%	91.50%	72.00%
Ours	97.00%	82.00%	93.50%	95.00%	93.50%	97.00%	83.00%	94.50%	95.00%	94.50%

According to Table 4, the proposed method performed well compared to other methods in indoor pomegranate cercospora spot and indoor and outdoor apple spotted leaf drop disease. For example, in the case of pomegranate cercospora spot, the segmentation accuracies of models such as HRNetV2, SegFormer, and U-Net in leaf and spot segmentation are far inferior to the proposed method. SegFormer performs best among the models compared, but not as well as the method proposed in this paper. In terms of segmentation mIoU for leaves and spots, the proposed method outperforms SegFormer by 2% and 9%, respectively. HRNetV2 is less accurate than the proposed method for target point segmentation. The mIoU for leaf segmentation was 1% lower, and for spot segmentation, it was 10% lower than the proposed method. PSPNet performed the worst in all comparative trials. The pyramid pooling module in PSPNet may be better in global contexts but is not suitable for agricultural diseases requiring both local and global feature attention. Statistical analysis shows that this method can effectively overcome the challenges of light problems, leaf curling, leaf overlapping, etc., that exist in real environments and can meet the requirements of detecting leaf diseases of apples, and pomegranate in real agricultural scenarios.

To better validate the performance of the proposed method in different scenarios, Table 5 shows the performance comparison of the method proposed in this paper and the comparison methods. Table 5 lists in detail the performance of all the methods when mIoU, mPA, total parameters, FLOPS, Acc, and F\_score are used as evaluation metrics, and, compared to other models, our proposed method performs well. DeepLabV3+ has shown the best performance among other models, but its segmentation accuracy is 10.78%, 9.51%, 0.72%, and 8.00% lower than the proposed methods mIoU, mPA, Acc, and F-score, respectively. PSPNet and U-Net have the worst overall segmentation performance, with a decrease of 28.48% and 19.86% in mIoU compared to this method. It is possible that their directionality is too single and they cannot adapt to agricultural disease tasks. In the other evaluation metrics, the proposed method does not achieve the best performance. SegFormer outperforms other methods in total parameters and FLOPS, while our method demonstrates a moderate performance, with total parameters and FLOPS being 2.85 M, 1.78 M, 70.08 G, and 0.32 G lower than U-Net and DeepLabV3+, respectively. Compared to PSPNet, our method has 2.95 M more total parameters but 9.81 G fewer FLOPS. The proposed method outperforms other models by approximately 10% on average, in terms of mIoU and mPA. Although it does not achieve the best results in total parameters and FLOPS, it ranks among the top, meeting the computational requirements for deployment on cloud servers. In summary, these results fully demonstrate the effectiveness and applicability of our method in the field of agricultural fruit leaf disease image segmentation, providing strong support and reference for future precision disease management in smart agriculture.

**Table 5.** Performance of different models for multi-fruit samples.

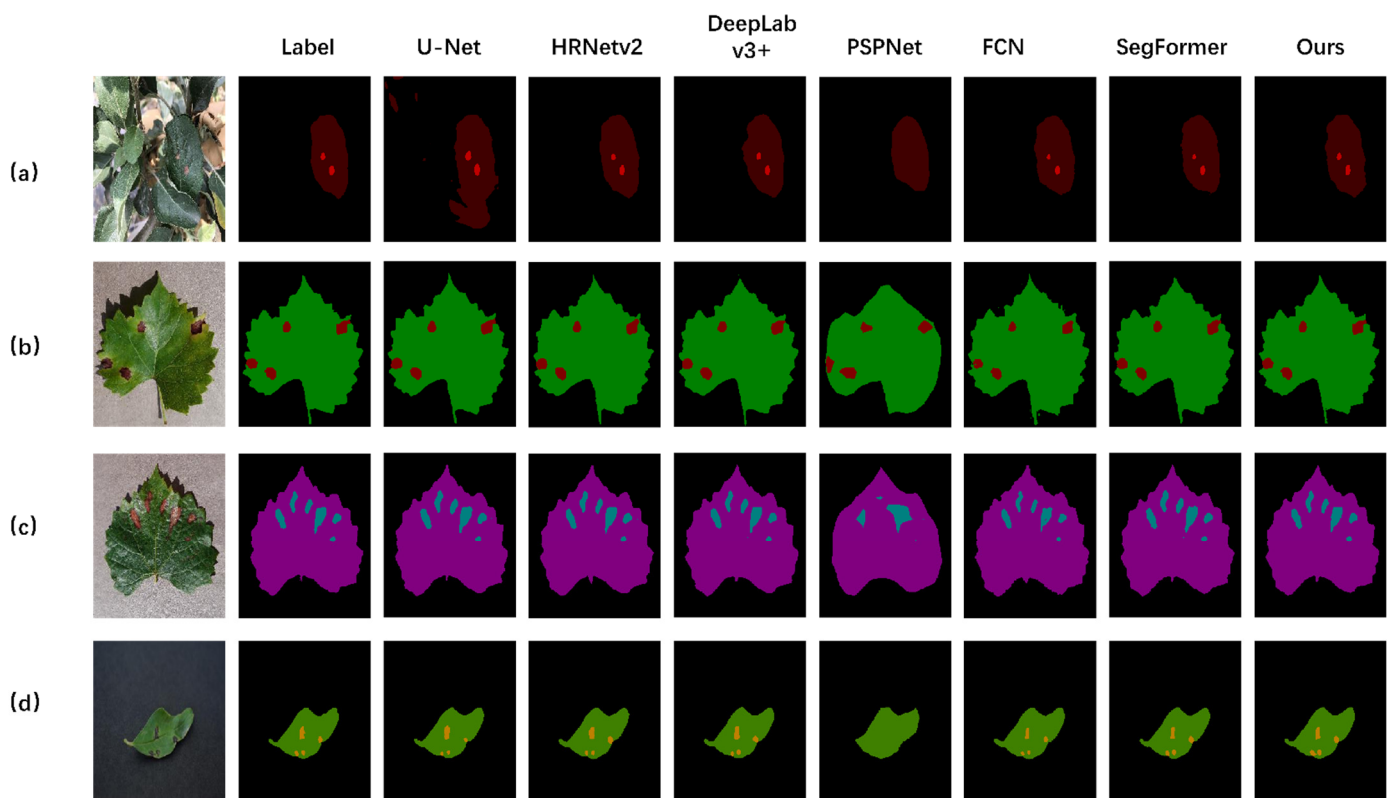
Method	mIoU	mPA	Acc	F_Score	Total Parameters/M	FLOPs/G
PSPNet	60.12%	65.83%	96.25%	70.80%	19.08	42.73
HRNetV2	79.04%	84.52%	98.10%	89.20%	22.77	29.54
U-Net	68.74%	75.67%	95.58%	79.60%	24.89	103.00
DeepLabv3+	77.82%	83.98%	97.88%	87.90%	23.81	33.24
SegFormer	77.35%	84.52%	96.25%	70.80%	3.72	3.41
FCN	70.49%	74.93%	89.11%	85.30%	18.88	30.23
Ours	88.60%	93.49%	98.60%	95.90%	22.03	32.92

Figure 7 shows the resultant images of four different diseases of three fruits using different models. The segmentation results from PSPNet are unsatisfactory, failing to clearly segment the edge parts and spots of each leaf. This is mostly due to the multi-scale pyramid pooling structure, which loses more details during downsampling, resulting in poor segmentation results. For example, PSPNet loses too much edge information in the case of the serrated blade of grape leaves, resulting in smooth leaf edges. In Figure 7a, apple spotted leaf drop disease is shown in a real environment. In the real environment with overlapping leaves and a complex background, the comparison of different methods shows the superiority of EFS-Former, which can solve the above problems by properly segmenting the shape of the leaves after shading, and accurately extracting the edge information of the diseased spots. In the case of U-Net, although U-Net effectively segments the spots compared to PSPNet, U-Net mistakenly segments the other leaves in the complex background of the real environment. In Figure 7b, the target is mainly changed to grape leaves affected by shadows only, and DeepLabV3+ and FCN are slightly improved compared to PSPNet, which can segment most of the jagged shapes of leaves but cannot accurately distinguish the edges of spots from the edges. As shown in Figure 7c, this method effectively segments the blurred leaf edges that are affected by shadows. SegFormer, which lacks local feature extraction, also fails to accurately segment the serrations of the leaf and is not precise enough to distinguish the edge contours of the target spots. In Figure 7d, all models except PSPNet can clearly segment the leaf blade. However, for diseased spot adherence, all comparative models lost this local information. The proposed method accurately extracts this information, demonstrating its ability to capture detail. In summary, the proposed method can accurately extract the fuzzy edge information of diseased spots, accurately segment the shape of diseased spots, and accurately calculate the area occupied by each target point, which can provide support for subsequent severity assessment work.

Figure 8 presents the segmentation results of different models on this study dataset.

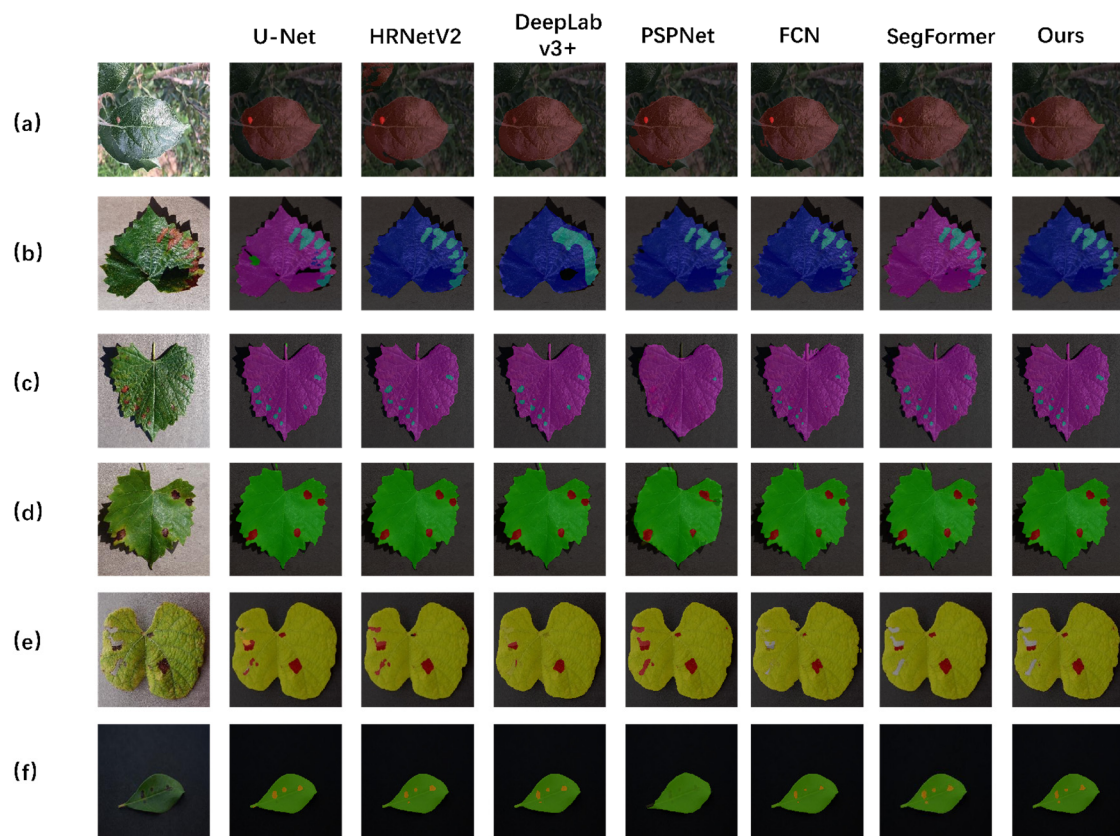
These include apple leaves in a real environment under strong light, which test the performance of the model in complex situations. Grape leaves at different disease periods verify the accuracy of the model in determining the disease period. Pomegranate leaves in a low-light environment verify the effectiveness of the model under different light conditions. Figure 8a shows apple spotted leaf drop disease in a complex field environment, subjected to bright light, weakening the spot features and causing confusion with the leaves, as well as diseased leaves adhering to other healthy leaves, thereby blurring the edge features of the leaves. Our method accurately segments the jagged edges of the leaves and the shape of the diseased spots, thus outperforming other methods. Models such as U-Net, HRNetV2, and SegFormer can accurately extract spots, but, in complex backgrounds, the segmentation of leaf edges is imprecise due to overlapping leaves and illumination, often mistaking healthy leaves for diseased ones. Deeplabv3+ is ineffective at extracting spots and fails to identify the features of apple spotted leaf drop disease in complex field environments. Figure 8b,c use different colors to represent the early and late disease stages of grape black rot. Our model accurately segments diseased leaf edges under light influence, overcomes the confusion between spot edges and leaves, accurately distinguishes diseased spots, and classifies the disease by the proportion of spot area. In

contrast, U-Net and SegFormer can segment roughly and clearly, but the edges of diseased spots are imprecise due to light influence and shading, leading to misclassification. U-Net mistakes shaded parts for diseased leaves and misclassifies late-stage disease as early-stage disease in segmentation. PSPNet, Deeplabv3+, and FCN can accurately locate the disease stage of grapes, but their segmentation of leaves and diseased spots is subpar. Light effects make the diseased spots similar in color to leaves, complicating feature extraction. In Figure 8d, U-Net, FCN, HRNetV2, and SegFormer struggle with leaf folding, especially under shadows, where the edges of diseased spots become blurred and cannot be accurately segmented. PSPNet poorly extracts leaves and diseased spots. In Figure 8e, the original leaf has rotting holes and disease spots, both indoors and outdoors, affecting segmentation and testing the model's feature extraction ability. FCN mistakenly treats the background as a leaf during segmentation, leading to inaccurate contextual feature extraction due to the restricted receptive field. In contrast, U-Net, Deeplabv3+, HRNetV2, and PSPNet can accurately segment leaf edges, but rot holes affect lesion extraction. SegFormer accurately extracts features of leaves and rot holes due to its global contextual advantage, but the lack of local features leads to detail loss. In Figure 8f, darker light blurs the spots. PSPNet fails to extract the spots despite accurately segmenting the leaves. U-Net, Deeplabv3+, and HRNetV2 miss smaller spots, while SegFormer segments each spot but loses more edge features. In summary, our method can still accurately capture local and global features under the influence of leaf folding, lighting changes, leaf edge shadows, complex backgrounds, etc. This is thanks to FFM's ability to integrate local and global contextual information, effectively overcoming real and simulated lighting conditions, achieving precise segmentation of leaves and lesions, and demonstrating strong robustness.



**Figure 7.** Images of results generated in different models for four disease conditions of three fruits: (a) apple spotted leaf drop disease, (b) grape brown spot disease, (c) grape black rot disease, and (d) pomegranate cercospora spot.

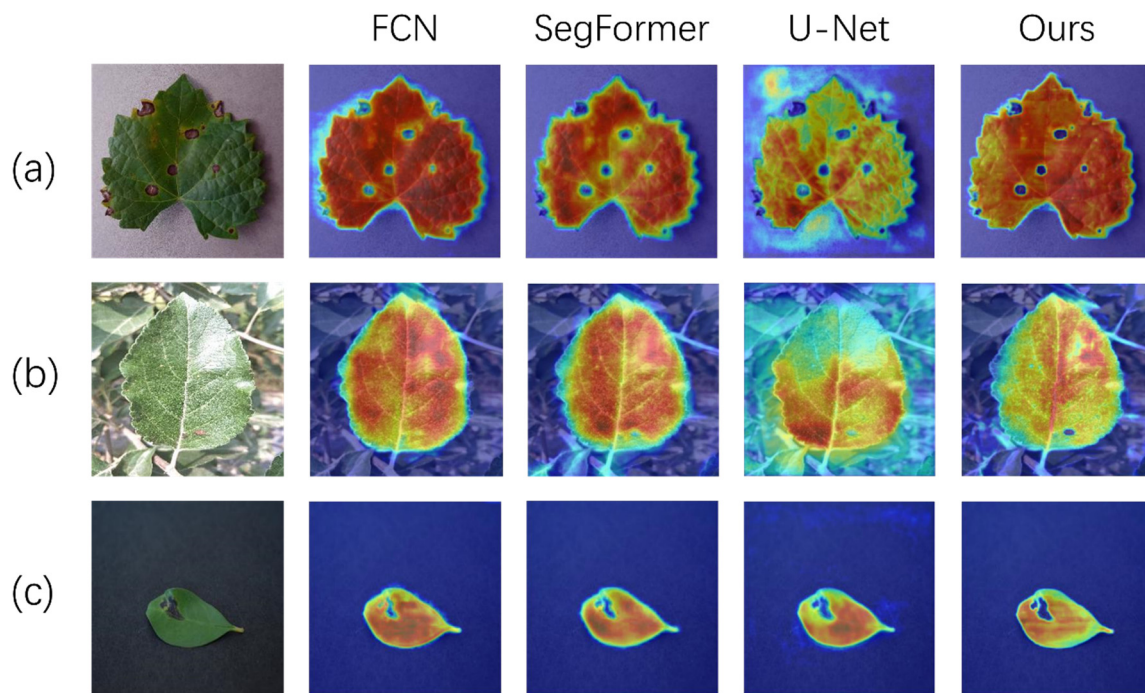




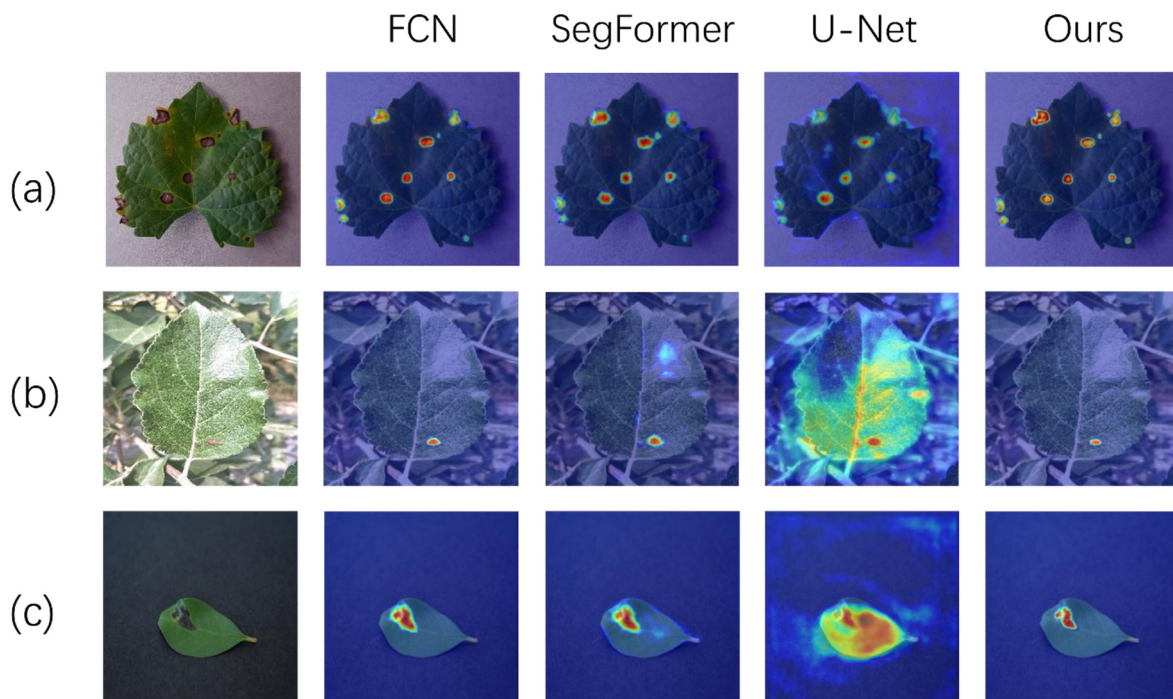
**Figure 8.** Visualization of segmentation results for different methods. (a) Apple spotted leaf drop disease, (b,c) late and early stages of grape black rot disease, (d,e) early and late stages of grape brown spot disease, and (f) pomegranate cercospora spot.

Figures 9 and 10 contain data samples that confuse the edge of diseased spots with those of the leaf, due to being subjected to different lighting conditions. To validate the accuracy of the feature extraction of diseased spots' edges in this paper, the effective regions of interest of the FCN, SegFormer, U-Net, and EFS-Former methods for the leaf and diseased spots were visualized, respectively. Figure 9a,b contain samples from both indoor and outdoor conditions, respectively, and, in (b), the blurring of the diseased spots and leaf features due to the excessive light definitely causes difficulties in feature extraction. Compared with other methods, EFS-Former has a stronger and more complete focus area on the overall focus area of the leaf; U-Net does not have a complete focus area in grape and apple leaves, mostly due to the influence of disturbances in the environment on U-Net, and therefore cannot accurately notice the target leaves. FCN, although it can efficiently focus on the overall portion of the leaf, confuses the leaf with the disease spots, and does not effectively delineate the diseased spot area when the target of attention is the leaf. For smaller pomegranate leaves in Figure 9c, EFS-Former demonstrates the ability to extract features with detailed, edge information that is focused on the leaf edges. In contrast, most of the other methods ignore the edges between the spots and the lesions. In Figure 10a,b, the focus of each method on the spots is mainly shown; FCN and SegFormer miss some lesion areas due to the high light intensity and the presence of pixels in the background similar to the color of the disease. In Figure 10c, other methods fail to accurately focus on the edge area of the diseased spots, which blurs the edge information of the spots, the method proposed in this paper can effectively focus on the blurred edge information and can accurately differentiate the characteristic information of the leaf and the diseased spots. In summary, the EFS-Former proposed in this paper can accurately focus on leaves and spots under different conditions. Because the ELD module and H-attention mechanism of the model accurately control the attention interval, it can effectively solve the problem of

blurring the edges of light blurred spots and fine spots in a real environment, and achieve better segmentation results.



**Figure 9.** Effective attention results of different models for indoor and outdoor leaves. (a) Grape brown spot disease, (b) apple spotted leaf drop disease, and (c) pomegranate cercospora spot.



**Figure 10.** Results of different models for effective attention to leaf spots indoors and outdoors. (a) Grape brown spot disease, (b) apple spotted leaf drop disease, and (c) pomegranate cercospora spot.

### 3.1.2. Different Attention Mechanisms

In this summary, we discuss the effect of different attention mechanisms on model performance and compare five popular attention mechanisms, namely CBAM Attention [36], SK Attention [37], SE Attention [38], CoAttention [39], and Global Attention [40], with the proposed H-attention. H-attention is compared and evaluated using mIoU, mPA, Acc, and F\_score metrics. The experimental results, shown in Table 6, indicate that our proposed H-attention achieves the best results, with 88.60% mIoU, 93.49% mPA, 98.60% Acc, and 95.90% F\_score. It enhances feature extraction for leaf and spot edges by superimposing different-sized convolutions, increasing the model's interest interval, and reducing computational redundancy. Thus, the model performs well in all evaluation metrics.

**Table 6.** The impact of different attention levels on model performance.

Method	mIoU	mPA	Acc	F_Score
CBAM	65.72%	70.52%	95.74%	88.40%
SK Attention	70.48%	72.87%	96.37%	90.40%
SE Attention	70.49%	74.93%	96.25%	89.40%
CoAttention	72.90%	76.25%	97.27%	90.80%
Global Context Attention	72.76%	76.28%	97.08%	90.60%
Multi-Head Attention	82.88%	96.98%	96.98%	90.90%
H-Attention	88.60%	93.49%	98.60%	95.90%

In comparison, SE and Global are slightly inferior to the proposed approach. Although the SE module enhances the model's representational ability by recalibrating the channels, and Multi-Head improves its attention capturing ability by processing multiple attention distributions in parallel, enhancing expressive ability and learning efficiency, the Global module improves perceptual ability by capturing remote dependencies. However, they still need to improve their ability to handle complex scenes and feature interactions compared to the H-attention module. CBAM has the lowest values in all evaluation metrics, probably due to unwanted deformations and distortions introduced during its transformation process, leading to degraded model performance. However, it is worth noting that CBAM significantly enhances feature perception at specific spatial and channel locations using channel attention and spatial attention, potentially performing better in tasks requiring specific spatial transformations.

Taken together, the good performance of H-attention is attributed to the fact that, by choosing different convolution kernel sizes, the model can adapt to different scales of features. The model also makes full use of the global information extraction ability of self-attention and the local feature extraction advantage of convolution to improve its ability to extract features of leaf and spot edges. Secondly, the selective convolution module reduces computational complexity through global average pooling, while maintaining attention to important features, thus showing good performance in all evaluation metrics.

### 3.2. Ablation Experiment

This section presents four sets of ablation experiments to validate the effectiveness of the proposed ELD module, Seg-Block, parallel fusion architecture, and feature fusion module for segmentation of leaf diseases in agricultural fruit trees. The overall idea is to gradually remove the modules from the proposed EFS-Former model. As shown in Table 7, Method 1 is the baseline model without any module, and Method 4 is the complete model proposed in this study. In Method 2, we first remove H-Attention from the proposed Seg-Block and conduct parallel experiments by combining the ELD module through the proposed parallel fusion architecture. The results of the four evaluation metrics, mIoU, Acc, F\_score, and mPA, show that, compared to Method 1, this experiment rises by 8.67%, 1.13%, 5.20%, and 7.11%, respectively. The results show that the ELD module and H-attention have good adaptability when improving segmentation accuracy and other aspects. In contrast to Method1, using separate Seg-Block and running serial architecture in Method 3

improves the metric scores by 8.09%, 0.99%, 4.90%, and 7.31%, respectively. It can be seen that our proposed attention mechanism brings the most substantial overall improvement in attention to relevant features in the data, confirming its key role in segmenting fruit and vegetable leaf diseases. Compared to the benchmark model, the synergistic effect of our proposed H-attention mechanism and other modules ultimately leads to peak model performance. EFS-Former shows significant improvements in evaluation metrics, especially in mIoU and mPA, which increased by 11.25% and 10.16%, respectively, emphasizing its role in refining segmentation granularity. Regardless of the design of the encoder, the introduction of the ELD module and H-attention can provide good evaluation metrics in the leaf disease segmentation task, which can effectively improve the performance of leaf and spot segmentation. A comparison of Method 4 with other schemes shows that the proposed method provides the best performance for leaf and lesion segmentation. The combined results show that using ELD and H-attention as single modules results in excellent performance, proving the effectiveness of the proposed model, and, in the case of the complete model compared to the single modules, both have improved. Therefore, all the improvements can significantly improve the segmentation performance of the leaf disease segmentation model.

**Table 7.** Impact of different modules on performance.

	H-Attention (Seg-Block)	Structure (FFM)	ELD	mIoU	Acc	F_Score	mPA
Method 1	-	-	-	77.35%	97.22%	88.60%	83.31%
Method 2	-	✓	✓	86.02%	98.35%	93.80%	90.42%
Method 3	✓	-	-	85.44%	98.21%	93.50%	90.62%
Method 4	✓	✓	✓	88.60%	98.60%	95.90%	93.49%

### 3.3. Disease Severity Assessment

To better illustrate the process of disease severity assessment, we list four fruit leaves with different severity levels, as shown in Table 8. Disease coverage can clearly reflect the degree of disease proliferation and provide an intuitive indicator for assessing disease severity, which can help growers achieve precise disease control. Disease ratio is calculated as follows:

$$Disease\ Ratio = \frac{S_{Disease}}{S_{Disease} + S_{Leaf}} \quad (26)$$

**Table 8.** Assessment of disease severity of different fruit leaves.




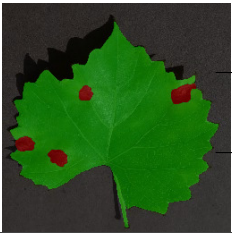




Original Image	Visualized Image	Label	Value	Ratio	Disease Ratio
		Background	214,067	81.66%	4.03%
		Leaf	46,137	17.60%	
		Disease spots	1940	0.74%	
		Background	143,367	54.69%	2.21%
		Leaf	113,849	43.43%	
		Disease spots	4928	1.88%	

Table 8. Cont.

Original Image	Visualized Image	Label	Value	Ratio	Disease Ratio
		Background	153,171	58.43%	3.22%
		Leaf	105,461	40.23%	
		Disease spots	3512	1.34%	
		Background	253,388	96.66%	5.98%
		Leaf	8231	3.14%	
		Disease spots	525	0.20%	

### 3.4. Discussion

The detection of leaf diseases in fruit trees is one of the commonly used methods for fruit disease detection, which can provide a practical basis for the prevention and control of early diseases. However, traditional leaf detection mainly relies on manual visual observation and long-term experience accumulation for judgment, and, for large orchards containing multiple fruits, this is not only time-consuming and laborious, but also limited in accuracy by the skill level of the detection personnel. Although deep learning methods have been widely used in leaf disease detection, they have the following shortcomings: (1) most leaf disease detection algorithms can only detect and locate diseases on leaves [41], but cannot evaluate the severity of diseases; (2) most algorithms based on leaf disease segmentation are limited to simple backgrounds or single leaf situations [42], ignoring the interference of complex backgrounds and multiple leaves in real scenes; (3) unclear leaf edges caused by leaf curling; and (4) the target disease spots are too small, which makes it easy to cause omissions when performing local feature extraction. The above limitations cause difficulties of most leaf disease detection algorithms when applying leaf disease detection in real scenarios.

This paper presents a parallel fusion model, fusing CNN and Transformer encoders, for the segmentation of leaf diseases in apple, grape, and pomegranate in complex environments. The CNN and Transformer branches in the model correspond to extracting local features and capturing global information, respectively, and fuse these two types of feature information through the FFM. The interference of complex background and multiple leaves in the real scene is effectively solved. In this paper, we propose the ELD module and H-attention mechanism. The ELD module expands the receptive field of the model by expanding convolution, and selective convolution in H-attention effectively focuses on the local region and reduces computational redundancy while extracting the global context, both of which effectively solve the problem of unclear leaf edges caused by the target diseased spots being too small and the leaves curling. Ultimately, the severity of fruit and vegetable diseases was determined by the area ratio of diseased spots to fruit tree leaf pixels. It can effectively help growers to accurately grasp the severity of the disease and realize the precise spraying of pesticides.

Some studies in plant leaf disease segmentation usually use a single CNN or Transformer architecture. Due to the different focus of these models, certain key features may be overlooked when dealing with different types of diseases and leaves. X. Zhang et al. [43] investigated the complementary nature of CNN and Transformer in image segmentation to simulate complex real-world environments by changing the original image background,

and achieved an IoU of more than 88.04% on the processed grape disease dataset. However, it is possible that most of the selected datasets are mainly from laboratory environments or that some of the details are lost in the fusion of global and local information, so they are prone to some fluctuations in the evaluation metrics. Jinhai Wang et al. [44] presented DualSeg, an image processing model based on two different branches of Swin Transformer and ResNet models, which segmented grapes in real vineyards with a mIoU of more than 83.7%. However, the model parameters are much higher than for the other models, probably due to the problem caused by the selection of a parallel encoder that is too large, which leads to difficulties in deployment. Also, the generalization of the model needs to be improved, since the whole experiment only involves one species, grapes, which is not applicable to large multi-species orchards. The results of these studies show that integrating the CNN and Transformer structures led to different degrees of accuracy improvement, in comparison to the above methods. This study achieved an overall improvement in the generalization of the model by targeting different species of multiple fruits from both laboratory environments and real conditions and reached 88.60% in terms of mIoU, which is significantly better than the single-crop study. At this stage, the main focus should be on improving the model's accuracy while enhancing the generalization of the model to fit the requirements in real-world environments.

Regarding its limitations, the EFS-Former implementation currently only concentrates on single-image processing. However, in real complex agricultural scenarios, it would be more practical to employ the continuity of surveillance video. The performance of the model on such real-time video streams, its ability to handle temporal coherence, and the computational challenges posed by this change remain to be explored. In this work, only three typical fruit diseased leaves were selected for the experiments, and the ability of EFS-Former to generalize to other fruit diseased leaves or other classes of fruit diseased leaves needs to be further validated and optimized. In future research work, we will continue to improve the EFS-Former model based on the deep learning tuning method, to improve the accuracy and reduce the overall computing cost. In addition, we will further investigate the feature samples of other kinds of fruit-like leaves and diseases, further test and validate the proposed EFS-Former in other fruit tree diseased leaf datasets and different environments, and extend it to diseased leaf segmentation and diagnosis of other crops to enhance the model's generalization ability. In summary, this work is an important attempt to apply the Transformer and CNN fusion network for leaf disease segmentation and diagnosis of fruit crops, and the methodology proposed in this paper provides a reference.

#### 4. Conclusions

The main challenge of this work is how to address issues like small disease spots, unclear leaf edges, and the confusion between disease spots and leaf colors under varying lighting conditions and occlusions in real-world environments. This study proposes the EFS-Former model, which uses a parallel fusion architecture capable of efficiently extracting both local and global features. The encoder utilizes both CNN and Transformer encoders to target local features and global context features, respectively. Additionally, the H-attention mechanism is introduced, to enhance the model's ability to capture global context information by obtaining the network's attention interval through convolutions of different sizes. This effectively improves the segmentation accuracy of disease spots, particularly when dealing with complex features like folded leaves or serrated edges, allowing for more precise extraction of disease spot boundaries and shapes, thereby increasing overall segmentation accuracy while reducing computational redundancy and complexity. By utilizing a lightweight ELD module as the encoder, this approach effectively expands the local receptive field, enhancing the ability to capture fine features and addressing the challenge of extracting edge features of leaves and spots. This enables the extraction of more micro-spots. The FFM effectively integrates two different types of features, reducing the loss of small disease spot information in leaves by combining various semantic features. The lightweight MLP (multilayer perceptron) layer, serving as the decoder, accurately

reconstructs the image by merging rich shallow features and edge information with high-level semantic features through multi-scale feature fusion. In terms of results, the proposed method achieved a mIoU of 88.60%, mPA of 93.49%, Acc of 98.60%, and an F<sub>score</sub> of 95.90%, representing an improvement of 11.25% and 10.16% in mIoU and mPA, respectively, compared to the baseline models. Additionally, this method accurately assessed the severity of three types of fruit leaf diseases by calculating the ratio of disease spot pixels to leaf area. Overall, the proposed EFS-Former demonstrates strong generalization and robustness, assisting farmers in the early detection of diseases through pathological image analysis of fruit leaves, enabling precise disease control and pesticide application while significantly reducing labor and material costs. In the future, we plan to conduct further experiments on a wide variety of diseased crop leaves to enhance the model's generalization capability. We also aim to adopt lightweight techniques and consider adjusting the model's depth to reduce the number of parameters. These strategies are intended to advance effective disease diagnosis across more crops and promote deployment on mobile platforms.

**Author Contributions:** D.J.: data curation, conceptualization, methodology, and writing—original draft. M.S.: conceptualization, methodology, and writing—review and editing. Z.Y.: data curation and writing—review and editing. S.L.: data curation and writing—review and editing. L.C.: writing—review and editing, supervision, and project administration. All authors have read and agreed to the published version of the manuscript.

**Funding:** This article was supported by the National Natural Science Foundation of China (Grant No. U19A2061), Agricultural image recognition and processing team, Jilin Provincial Science and Technology Department of young and middle-aged scientific and technological innovation and entrepreneurship excellence talent (team) project (innovation category) (Grant No. 20220508133RC), Jilin Province Science and Technology Development Plan Project (Grant No. 20210404020NC).

**Data Availability Statement:** The data that support the findings of this study are available from the corresponding author upon reasonable request.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

CNN	Convolutional neural network
ELD	Expanded local detail
FFM	Feature fusion module
MLP	Multilayer perceptron
FNN	Feedforward neural network
mIoU	Mean intersection over union
Acc	Accuracy
mPA	Mean pixel accuracy
FCN	Fully convolutional network

## References

- Shahi, T.B.; Xu, C.-Y.; Neupane, A.; Guo, W. Recent Advances in Crop Disease Detection Using Uav and Deep Learning Techniques. *Remote Sens.* **2023**, *15*, 2450. [[CrossRef](#)]
- Chin, R.; Catal, C.; Kassahun, A. Plant Disease Detection Using Drones in Precision Agriculture. *Precis. Agric.* **2023**, *24*, 1663–1682. [[CrossRef](#)]
- Coulibaly, S.; Kamsu-Foguem, B.; Kamissoko, D.; Traore, D. Deep Learning for Precision Agriculture: A Bibliometric Analysis. *Intell. Syst. Appl.* **2022**, *16*, 200102. [[CrossRef](#)]
- Febrinanto, F.G.; Dewi, C.; Triwiratno, A. The Implementation of K-Means Algorithm as Image Segmenting Method in Identifying the Citrus Leaves Disease. *IOP Conf. Ser. Earth Environ. Sci.* **2019**, *243*, 012024. [[CrossRef](#)]
- Chodey, M.D.; Shariff, C.N. Pest Detection Via Hybrid Classification Model with Fuzzy C-Means Segmentation and Proposed Texture Feature. *Biomed. Signal Process. Control.* **2023**, *84*, 104710. [[CrossRef](#)]
- Jothiaruna, N.; Sundar, K.J.A.; Karthikeyan, B. A Segmentation Method for Disease Spot Images Incorporating Chrominance in Comprehensive Color Feature and Region Growing. *Comput. Electron. Agric.* **2019**, *165*, 104934. [[CrossRef](#)]
- Ma, J.; Du, K.; Zhang, L.; Zheng, F.; Chu, J.; Sun, Z. A Segmentation Method for Greenhouse Vegetable Foliar Disease Spots Images Using Color Information and Region Growing. *Comput. Electron. Agric.* **2017**, *142*, 110–117. [[CrossRef](#)]

8. Chen, C.; Wang, X.; Heidari, A.A.; Yu, H.; Chen, H. Multi-Threshold Image Segmentation of Maize Diseases Based on Elite Comprehensive Particle Swarm Optimization and Otsu. *Front. Plant Sci.* **2021**, *12*, 789911. [[CrossRef](#)] [[PubMed](#)]
9. Xie, Z.; Wu, J.; Tang, W.; Liu, Y. Advancing Image Segmentation with Dbo-Otsu: Addressing Rubber Tree Diseases through Enhanced Threshold Techniques. *PLoS ONE* **2024**, *19*, e0297284. [[CrossRef](#)]
10. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
11. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015.
12. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
13. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected Crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [[CrossRef](#)]
14. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [[CrossRef](#)] [[PubMed](#)]
15. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-Cnn. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
16. Jia, W.; Tian, Y.; Luo, R.; Zhang, Z.; Lian, J.; Zheng, Y. Detection and Segmentation of Overlapped Fruits Based on Optimized Mask R-Cnn Application in Apple Harvesting Robot. *Comput. Electron. Agric.* **2020**, *172*, 105380. [[CrossRef](#)]
17. Zou, K.; Chen, X.; Wang, Y.; Zhang, C.; Zhang, F. A Modified U-Net with a Specific Data Argumentation Method for Semantic Segmentation of Weed Images in the Field. *Comput. Electron. Agric.* **2021**, *187*, 106242. [[CrossRef](#)]
18. Kang, J.; Liu, L.; Zhang, F.; Shen, C.; Wang, N.; Shao, L. Semantic Segmentation Model of Cotton Roots in-Situ Image Based on Attention Mechanism. *Comput. Electron. Agric.* **2021**, *189*, 106370. [[CrossRef](#)]
19. Azizi, A.; Abbaspour-Gilandeh, Y.; Vannier, E.; Dusséaux, R.; Msere-Gundoshmian, T.; Moghaddam, H.A. Semantic Segmentation: A Modern Approach for Identifying Soil Clods in Precision Farming. *Biosyst. Eng.* **2020**, *196*, 172–182. [[CrossRef](#)]
20. Wang, C.; Du, P.; Wu, H.; Li, J.; Zhao, C.; Zhu, H. A Cucumber Leaf Disease Severity Classification Method Based on the Fusion of Deeplabv3+ and U-Net. *Comput. Electron. Agric.* **2021**, *189*, 106373. [[CrossRef](#)]
21. Sunil, C.K.; Jaidhar, C.D.; Patil, N. Tomato Plant Disease Classification Using Multilevel Feature Fusion with Adaptive Channel Spatial and Pixel Attention Mechanism. *Expert Syst. Appl.* **2023**, *228*, 120381.
22. Liu, B.-Y.; Fan, K.-J.; Su, W.-H.; Peng, Y. Two-Stage Convolutional Neural Networks for Diagnosing the Severity of Alternaria Leaf Blotch Disease of the Apple Tree. *Remote Sens.* **2022**, *14*, 2519. [[CrossRef](#)]
23. Dai, G.; Fan, J.; Tian, Z.; Wang, C. Pplc-Net: Neural Network-Based Plant Disease Identification Model Supported by Weather Data Augmentation and Multi-Level Attention Mechanism. *J. King Saud Univ. Comput. Inf. Sci.* **2023**, *35*, 101555. [[CrossRef](#)]
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
26. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-Local Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
27. Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-Attention Generative Adversarial Networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019.
28. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
29. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.S. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021.
30. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. Segformer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
31. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR), Nashville, TN, USA, 20–25 June 2021.
32. Hughes, D.; Salathé, M. An Open Access Repository of Images on Plant Health to Enable the Development of Mobile Disease Diagnostics. *arXiv* **2015**, arXiv:1511.08060.
33. Russell, B.C.; Torralba, A.; Murphy, K.P.; Freeman, W.T. LabelMe: A Database and Web-Based Tool for Image Annotation. *Int. J. Comput. Vis.* **2008**, *77*, 157–173. [[CrossRef](#)]
34. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3349–3364. [[CrossRef](#)] [[PubMed](#)]



35. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
36. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
37. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective Kernel Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
38. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
39. Lu, J.; Yang, J.; Batra, D.; Parikh, D. Hierarchical Question-Image Co-Attention for Visual Question Answering. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 289–297.
40. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. Gcnet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October 2019.
41. Ahmed, K.; Shahidi, T.R.; Alam, S.M.I.; Momen, S. Rice Leaf Disease Detection Using Machine Learning Techniques. In Proceedings of the 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh, 24–25 December 2019.
42. Wang, Y.; Wang, H.; Peng, Z. Rice Diseases Detection and Classification Using Attention Based Neural Network and Bayesian Optimization. *Expert Syst. Appl.* **2021**, *178*, 114770. [[CrossRef](#)]
43. Zhang, X.; Li, F.; Zheng, H.; Mu, W. Upformer: U-Sharped Perception Lightweight Transformer for Segmentation of Field Grape Leaf Diseases. *Expert Syst. Appl.* **2024**, *249*, 123546. [[CrossRef](#)]
44. Wang, J.; Zhang, Z.; Luo, L.; Wei, H.; Wang, W.; Chen, M.; Luo, S. Dualseg: Fusing Transformer and Cnn Structure for Image Segmentation in Complex Vineyard Environment. *Comput. Electron. Agric.* **2023**, *206*, 107682. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.