

Article

Plant Diseased Lesion Image Segmentation and Recognition Based on Improved Multi-Scale Attention Net

Tao Yang, Yannian Wang and Jihong Lian *

School of Electronic Information, Graduate School of Xi'an Engineering University, Lintong Campus, Xi'an 710000, China; 210421094@stu.xpu.edu.cn (T.Y.); wyn11405@163.com (Y.W.)

* Correspondence: lianjihong@163.com; Tel.: +86-185-02986391

Abstract: Fallen leaf disease can lead to a decrease in leaf area, a decrease in photosynthetic products, insufficient accumulation of fruit sugar, poor coloring and flavor, and a large number of fruits developing sunburn. To address the aforementioned issue, this article introduces a deep learning algorithm designed for the segmentation and recognition of agricultural disease images, particularly those involving leaf lesions. The essence of this algorithm lies in enhancing the Multi-scale Attention Net (MA-Net) encoder and attention mechanism to improve the model's performance when processing agricultural disease images. Firstly, an analysis was conducted on MA-Net, and its limitations were identified. Compared to res-block, Mix Vision Transformer (MiT) consumes relatively less time during the training process, can better capture global and contextual information in images, and has better robustness and scalability. Then, the feature extraction parts of different networks were used as encoders to join the MA-Net network. Compared to a Position-wise Attention Block (PAB), which has higher computational complexity and requires a larger amount of computing resources, Effective Channel Attention net (ECANet) reduces the number of model parameters and computation by learning the correlation between channels, as well as having a better denoising ability. The experimental results show that the proposed solution has high accuracy and stability in agricultural disease image segmentation and recognition. The mean Intersection over Union (mIoU) is 98.1%, which is 0.2% higher than traditional MA-Net; Dice Loss is 0.9%, which is 0.1% lower than traditional MA-Net.

Keywords: leaf lesions segmentation; agriculture; deep learning; accurate management; attention; feature extraction; computer vision



Citation: Yang, T.; Wang, Y.; Lian, J. Plant Diseased Lesion Image Segmentation and Recognition Based on Improved Multi-Scale Attention Net. *Appl. Sci.* **2024**, *14*, 1716. <https://doi.org/10.3390/app14051716>

Academic Editor: Sungho Kim

Received: 11 January 2024

Revised: 5 February 2024

Accepted: 12 February 2024

Published: 20 February 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

China is a major agricultural producer, and agriculture is the foundation of a powerful country. Agricultural issues are related to people's livelihoods and are one of the fundamental issues in China. In the global agricultural system, China has always been a major export country. Crop disease is one of the main agricultural disasters in China, characterized by multiple types, significant impacts, and frequent outbreaks [1]. Its scope and severity often cause significant losses to the national economy, especially agricultural production.

There are several examples of apple leaf lesions. (1) *Alternaria* blotch mainly damages leaves, petioles, annual branches, and fruits. It is one of the important diseases that causes early leaf drop in apples, resulting in weakened tree vigor and reduced fruit yield and quality. (2) Brown spot mainly damages leaves, followed by fruits and petioles. It is one of the most important diseases, causing early defoliation of apples, and it is also the most common and serious disease which causes defoliation of apple trees. (3) Grey spot mainly damages leaves, but also branches, petioles, shoots, and fruits. The affected leaves generally do not turn yellow and fall off, but severely affected leaves may appear scorched. (4) Mosaic mainly damages leaves, causing discoloration, ring rot, distortion, and shrinkage of leaves on severely affected trees. The disease slows down the growth of affected trees, sometimes

leading to reduced yields and quality of early leaves. Affected fruits are not durable in storage and are prone to inducing other diseases. (5) Rust mainly damages leaves, but can also affect young branches, young fruits, and fruit stems, often causing early defoliation, weakening tree vigor, and affecting yield.

The hazards of early fallen leaves mainly include the following aspects: a decrease in leaf area leads to a decrease in photosynthetic products, which in turn affects the accumulation of sugar in the fruit, resulting in color differences and poor flavor. If there is no leaf cover, a large amount of fruit will be sunburned, further affecting the quality of the fruit. Insufficient nutrition in the flower buds can lead to poor differentiation or degeneration, and there may be no flowers or fruits the following year. The storage material of fruit trees decreases in winter, and the leaves and branches are thin and weak in spring, making them susceptible to cold and freezing damage. The early shedding of leaves can easily lead to secondary flowering, further affecting the yield of flowers and fruits the following year. The early shedding of leaves can seriously affect the vitality of trees, making them more susceptible to other diseases and posing a threat to the health of fruit trees. Disease problems have become a major factor limiting the growth of agricultural production. As agricultural production becomes increasingly dependent on refined management, manual disease treatment has become very cumbersome and unsustainable.

In recent years, deep learning technology has been applied in many fields, especially in computer vision. Deep learning's effectiveness in image segmentation has been widely researched and confirmed. In the agricultural field, deep learning can be used for agricultural image segmentation problems, such as distinguishing disease areas in disease detection. This article will introduce the application of deep learning in agricultural disease spot image segmentation. Firstly, the article will discuss the relationship between disease detection and agricultural production, as well as the challenges they present. Next, it will explain the principles of deep learning and the basic concepts of image segmentation. Then, it will discuss how to apply deep learning to agricultural disease spot image segmentation and verify its accuracy and effectiveness through experiments. Finally, the potential value and future development direction of deep learning in agricultural production will be explored.

Contribution of this article:

- (1) We annotated the original lesion images and split the dataset into five common apple leaf pathology images, including alternaria blotch, brown spot, grey spot, mosaic, and rust.
- (2) We propose an improved Multi-scale Attention Net (MA-Net) model using a Mix Vision Transformer (MiT) as the encoder to capture global and contextual information in images and using Effective Channel Attention (ECA) instead of MA-Net's Position-wise Attention Block (PAB) module to reduce the number of model parameters and computation by learning the correlation between channels.
- (3) Through comparative experiments, it has been proven that the improved model achieves good segmentation results on the apple leaf lesion dataset, with Dice Loss of 0.9%, mean Intersection over Union (mIoU) of 98.1%, and accuracy of 99.6%.

The remainder of this paper is arranged as follows: in Section 2, we summarize previous research findings and introduce traditional image segmentation methods and the application of neural networks in the field of images; Section 3 is the methodology, which introduces the improvement methods for MA-Net, production of the dataset, and the experimental setup; Section 4 is the results, where we describe the experimental results; in Section 5 we analyze the research results and explain the superiority of the method; Section 6 is the Conclusion, where we summarize the experimental results, discuss limitations, and propose future research directions.

2. Related Works

Early disease and insect image recognition is mainly based on traditional machine learning, which uses image preprocessing (grayscale, histogram equalization, color space conversion, median filtering, etc.), image segmentation, and feature extraction techniques (lo-

cal binarization, grayscale co-occurrence matrix, color co-occurrence matrix, inter grayscale correlation matrix, etc.) to manually select and design features. Afterwards, specific machine learning algorithms such as support vector machines, random forests, linear regression analysis, principal component analysis, etc., are used to train their classification ability for specific feature vectors in order to achieve the goal of identifying pest and disease images [2].

Tian, Y. et al. [3] trained and optimized the SVM model, by extracting the color, shape, and texture features of grape disease spots, to identify grape diseases. They found that the comprehensive features of color and texture have stronger expression ability than single features. Li, Z. et al. [4] used the Back Propagation (BP) neural network to recognize and classify the shape, color, and texture features of extracted grape disease spots; their results showed that the recognition accuracy of grape disease spot samples reached 92.6%. In 2015, Kaur, R. [5] conducted a study on the use of SVM to identify crop diseases. In order to obtain clearer features, this method uses image segmentation algorithms to separate leaves from the background. After adjusting the color tone, SVM can be used to classify diseases. Huang, S. et al. [6] constructed a bag of words model for rice diseases using multispectral images and identified the detected rice diseases using a bag of words dictionary. In the end, they achieved a recognition accuracy of 94.72%, far higher than traditional spectral image analysis methods. Podol et al. [7] utilized the K-Means clustering algorithm to preprocess the identified samples and separate diseased grape leaves from the background. SVM was used to extract disease classification from leaves, achieving a recognition rate of 88.89% for grape diseases. Ma, Y. et al. [8] used object detection methods to determine the position of grape leaves, extracted features of grape diseases using HOG (directional gradient histogram), and then used SVM for classification and recognition. Islam M [9] identified potato diseases based on multi-classification SVM. After separating the background and lesions from the diseased leaves, they extracted the color and texture features of potato lesions. Finally, using multi-class support vector machines to classify the extracted features, they achieved a recognition rate of 95% for potato diseases. In 2018, Wang, Z. et al. [10] proposed a method for identifying rice blast spores using additive cross kernel support vector machines based on HOG features. This solved the shortcomings of traditional recognition methods such as a cumbersome process, strong subjective dependence, and low recognition rate, and achieved automatic recognition of rice blast spores. In 2019, Sun et al. [11] used a simple linear iterative clustering algorithm as a preprocessing step for segmentation to extract the significance map of tea plant diseases. They extracted superpixel feature blocks from multiple directions using a grayscale co-occurrence matrix, and used a support vector machine to classify and detect tea diseases. Zhu, R. et al. [12] used denoising filtering algorithms to denoise the original image of barley leaves, and separated the disease spots in the barley leaves from the background based on the color and texture differences of different diseases. Support vector machines were then used to identify the texture and color characteristics of barley disease spots.

Deep learning is a popular research area in artificial intelligence, particularly prevalent in the field of computer vision. Deep learning, a branch of machine learning, is a multi-processing layer computational model composed of multiple nonlinear transformations, and combines shallow features of data to represent abstract high-dimensional semantic information. In 2006, Hinton et al. [13], a leading figure in machine learning, first proposed the concept of deep learning in the renowned academic journal *Science*. They used self-supervised learning algorithms of deep trust networks to initialize network weights, solving the problem of difficulties in optimizing deep learning models and triggering a new wave of deep learning research.

Sladojevic et al. [14] proposed a plant disease recognition algorithm based on fine tuning deep convolutional neural networks capable of detecting plant leaves and distinguishing between 13 different plant diseases and healthy leaves, achieving a recognition accuracy of 96.3%. Brahimi et al. [15] proposed a tomato disease image recognition algorithm based on deep convolutional neural networks. CNN was used to extract features from tomato disease images, and visualization was used to locate tomato disease areas, achieving

high recognition accuracy. Xiao, Z. et al. [16] used the Hough transform to locate potato leaf disease areas and used morphological segmentation to segment disease areas. They also used principal component analysis to adaptively fuse the color and texture features of the disease, achieving rapid recognition of potato diseases. Veeraballi et al. [17] proposed a deep learning based papaya disease image recognition algorithm which classifies papaya disease images based on ResNet50. The algorithm can accurately identify papaya diseases under complex conditions such as insufficient lighting and different image sizes, and has high robustness. Zhang et al. [18] proposed an enhanced corn disease image recognition algorithm using GoogLeNet. By adjusting parameters, improving pooling layers, and reducing the number of classifiers, they improved both recognition accuracy and model training efficiency for corn leaf diseases. Silva et al. [19] proposed a plant disease image recognition algorithm based on multichannel convolutional neural networks. The algorithm did not introduce pretraining parameters, but instead used the PlantVillage dataset to train multichannel CNN from scratch. The model was trained on a segmented RGB grayscale RGB dataset, which accelerated the training speed and significantly improved the recognition accuracy of plant disease images. Hu, Z. et al. [20] extracted dispersed early disease features of tomato leaves through a multi-layer attention residual module. To enhance feature reuse, they fused low-level and high-level features to achieve fine-grained recognition of tomato leaf diseases. Sharma et al. [21] used FCN to segment tomato leaf diseases, and then used convolutional neural networks to recognize the segmented lesion images. The recognition accuracy was significantly improved, but the recognition effect was poor for diseases with dense lesions. Jiang et al. [22] proposed a rice leaf disease recognition algorithm based on an improved convolutional neural network. The algorithm combines CNN with support vector machine to extract features from rice leaf disease images. The features extracted by CNN are classified using support vector machine. The optimal parameters of the support vector machine are obtained through tenfold cross-validation, and the accuracy on the rice leaf disease image test set is 96.8%. Hou Jinxiu et al. [23] designed a residual attention network to extract multichannel features, recalibrated them, and fused them to solve the problem of identifying multiple types of plant diseases. Su et al. [24] studied a DNN algorithm for real-time segmentation of weeds between rows. Compared with DNN using traditional encoders, two new subnet structures were adopted to improve segmentation accuracy. Liu, Y. et al. [25] used U-Net to perform semantic segmentation on corn leaves and four common corn diseases, achieving better segmentation results than traditional image segmentation methods. Zhong Changyuan et al. [26] proposed an attention module based on group activation strategy, which uses high-order features to guide the enhancement of low-level features. By calculating the intra group enhancement coefficient, the suppression effect between different groups is reduced, and the feature expression ability is enhanced. ResNet18 is used to extract features, completing the segmentation of six common disease images of cucumber and rice and improving the segmentation accuracy. Su et al. [27] introduced a new data augmentation framework based on random cropping and RICAP. This framework is utilized to enhance image classification data and expand semantic segmentation tasks, resulting in improved segmentation accuracy. Jiang et al. [28] proposed a multitask recognition algorithm based on improved VGG16. The algorithm uses transfer learning to fine tune the parameters of VGG16 and trains the model using alternating learning. It can simultaneously recognize wheat leaf diseases and rice leaf diseases. Compared with the single task method, the proposed multitask recognition algorithm has fewer training parameters and higher recognition accuracy. Xiangpeng, F. et al. [29] proposed an improved CNN based corn disease recognition algorithm and designed a convolutional neural network structure with five layers of convolution, four layers of pooling, and two fully connected layers. The network was optimized using L2 regularization and Dropout strategy and verified on a corn disease image test set. The average recognition accuracy was 97.10%. Su, S. et al. [30] transferred the knowledge learned from ImageNet to VGG16 to accelerate model training speed and achieve small sample recognition of grape leaf diseases. Li, Q. et al. [31] transferred pretrained weights from

ImageNet to ResNet, and designed an attention residual module to reduce the number of parameters, solving the problem of rapid identification of corn diseases with small amounts of data. Zhang, N. et al. [32] designed a multi-scale convolutional module based on an attention mechanism to improve the ability of effective feature extraction, and applied it to InceptionV3. At the same time, transfer learning was used during the training process to avoid overfitting, achieving tomato leaf disease recognition. Wan, J. et al. [33] used transfer learning and GoogLeNet to achieve fruit tree disease identification and severity grading. Bao, W. et al. [34] introduced selective convolution in VGG16 to extract small disease features and used transfer learning to train the model, solving the problem of identifying small apple diseases.

3. Methodology

3.1. Proposed Model

MA-Net is an architecture based on U-Net with an added self-attention mechanism for image segmentation. This mechanism captures the spatial and channel dependencies of feature maps, taking into account multi-scale semantic information based on channel dependencies between feature maps. As shown in Figure 1, we replaced the Res-blocks of traditional MA-Net with MiT encoders as the encoder for the model. Replacing the PAB of the original model with ECA. Connect the corresponding layers in the encoder and decoder through skip connections, and enhance useful feature maps and suppress feature maps with smaller contributions through MFAB.

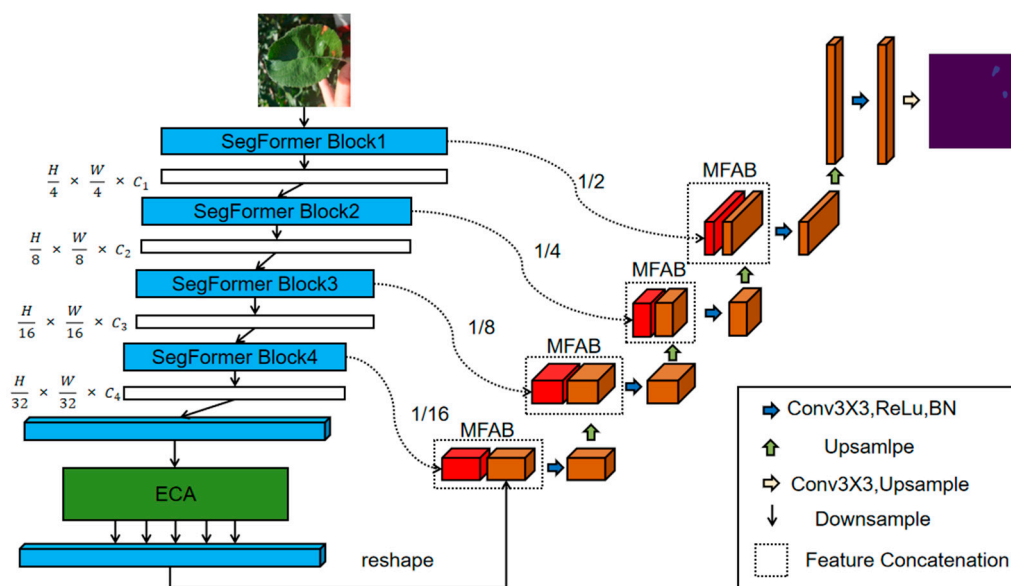


Figure 1. The proposed structure of improved model.

The traditional MA-Net network structure mainly consists of three modules: Resblock, Position-wise Attention Block (PAB), and Multi-scale Fusion Attention Block (MFAB). The improved MA-Net network in this article uses the Mix Transformer encoder (MiT) for mit_b2 which replaces Res-block as the encoder for the new model. Using ECA attention modules instead of PAB allowed us to capture spatial dependencies between any two positional feature maps and model a wider range of rich spatial contextual information on local feature maps.

3.2. Mix Vision Transformer

The backbone comes from SegFormer [35], a simple, efficient, powerful semantic segmentation framework that unifies a Transformer with lightweight Multilayer Perceptron (MLP) decoders. Main features: (1) a new hierarchical Transformer encoder that outputs

multi-scale features without positional encoding. This avoids interpolation of positional codes, reducing performance degradation when testing resolution differs from training. (2) The MLP decoder avoids complexity by aggregating information from different layers, combining local and global attention and offering strong representation capabilities. As shown in Figure 2, the overall network model consists of two parts: an encoder and a decoder. The encoder is composed of a layered Transformer module, while the decoder is composed of MLP.

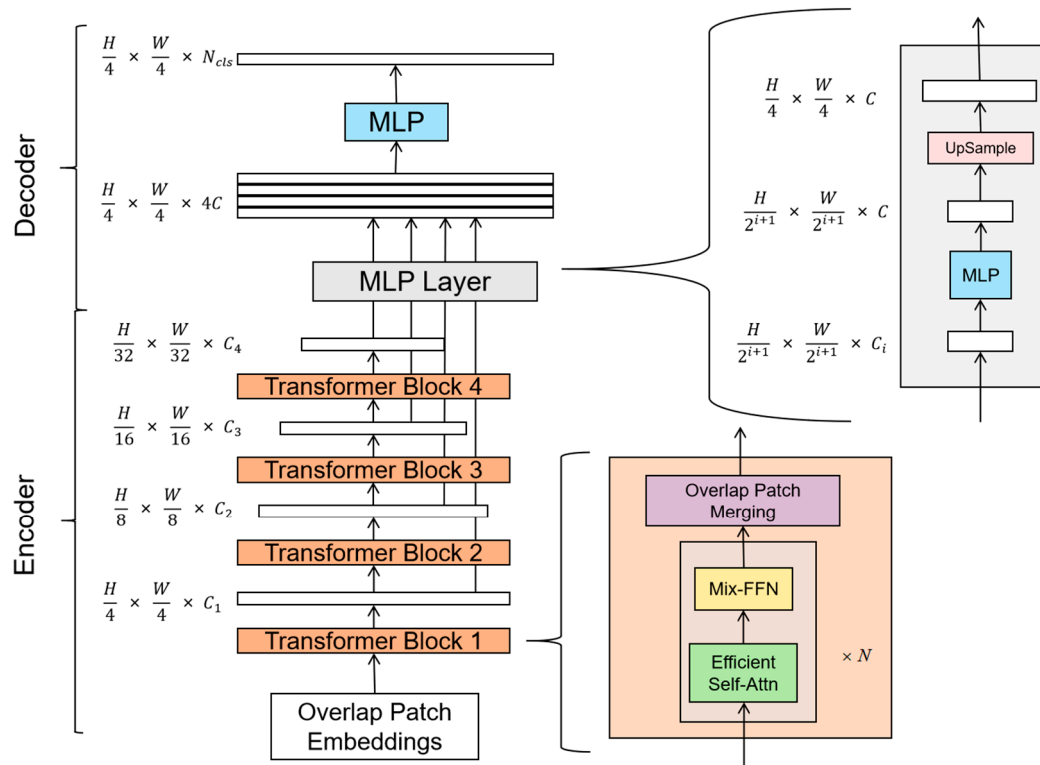


Figure 2. The SegFormer network has two main modules: a hierarchical Transformer encoder to extract features, and a lightweight All-MLP decoder to fuse these features and predict the segmentation mask. The FFN stands for feed-forward network [35].

The Transformer first achieved success in the natural language processing (NLP) field, and later Dosovitskiy et al. pioneered the use of Vision Transformer (ViT) in the CV field, achieving comparable results to CNN. ViT divides an image into fixed-sized patches, linearly embeds each patch, adds positional embeddings, and provides the resulting vector sequence to a standard Transformer encoder. For classification purposes, ViT uses a standard method of adding additional learnable “classification markers” to the sequence.

The standard Transformer receives a one-dimensional sequence embedded with a flag as input. To process two-dimensional images, ViT reshapes $x \in R^{H \times W \times C}$ into a series of flat two-dimensional patches where $x_p \in R^{N \times (P^2 \times C)}$, (H, W) represents the resolution of the original image, C represents the number of channels, (P, P) represents the resolution of every image patch, and $N = \frac{H \times W}{P^2}$ represents the number of patches, and can also be used as an effective input sequence length for the Transformer. The Transformer uses a constant latent vector size D in all layers, so ViT flattens the patch and maps it to the D dimension through trainable linear projection. The output of ViT projection is called patch embedding.

SegFormer is a cutting-edge semantic segmentation Transformer framework that combines efficiency, accuracy, and robustness. Compared to traditional methods, SegFormer’s framework has redesigned the encoder and decoder. Firstly, the encoder does not use interpolation position codes when inferring images of different resolutions from the training images. Therefore, SegFormer’s encoder can easily adapt to any testing resolution

without affecting performance. Furthermore, the layered part allows the encoder to generate low-resolution rough features and high-resolution fine features, which contrasts with ViT's ability to generate a single descending feature map with a fixed resolution. Secondly, SegFormer proposed a lightweight MLP decoder that utilizes Transformer induced features, where lower-level attention tends to maintain locality, while higher-level attention is highly non local. By aggregating information from different layers, local and global attention is combined by the MLP decoder. Therefore, SegFormer has obtained a concise and direct decoder that can demonstrate strong expressive power. SegFormer has created a range of Mix Vision Transformer (MiT) encoders, from mit-b0 to mit-b5, with identical structures but varying sizes.

3.3. ECA

Effective Channel Attention (ECA) is an improved version of SENet, which removes the fully connected layer from the original SENet and replaces it with a 1×1 convolutional kernel for processing, making the model parameters smaller and lighter.

As shown in Figure 3, obtain aggregated features of input features through Global Average Pooling (GAP), and ECA generates channel weights by performing fast 1D convolution with a size of k . Among them, k is adaptively determined by mapping the channel dimension C . The goal of the ECA module is to appropriately capture local interactions across channels, so it is necessary to determine the coverage range of interactions (i.e., the kernel size k of one-dimensional convolution). In different CNN architectures, the optimized interaction coverage can be manually adjusted for convolutional blocks with different channel numbers. However, manual tuning through cross validation will consume a significant amount of computing resources.

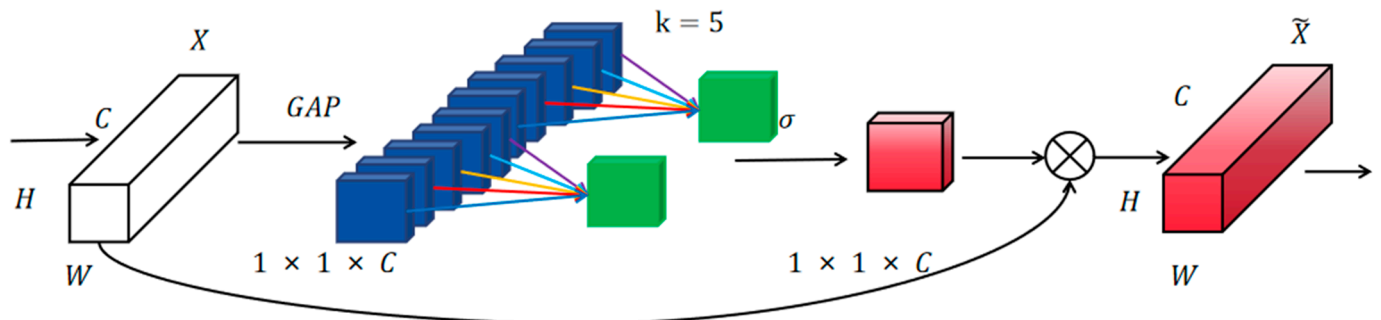


Figure 3. The structure of ECA model [36].

Therefore, a possible mapping was introduced:

$$C = \phi(k) = 2^{(\gamma * k - b)} \quad (1)$$

In this paper, we set γ and b to 2 and 1 for all experiments.

The kernel size k can be adaptively determined based on the channel dimension C :

$$k = \psi(C) = \left\lceil \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right\rceil_{\text{odd}} \quad (2)$$

3.4. MFAB

The MFAB module aims to learn the importance of each feature channel from multi-level feature maps without considering additional spatial dimensions. It can strengthen useful feature maps based on the importance of each channel and weaken feature maps that contribute less to the segmentation task. Specifically, it can describe the interdependence between feature channels in high-level and low-level feature maps. Advanced features have rich image semantic information, while low-level features that skip connections have more edge information. Low-level features are used to restore the details of an image.

Applying channel attention mechanisms to high-level and low-level features separately enabled us to increase the weight of key information in each channel while reducing the impact of irrelevant information.

As shown in Figure 4, we first input the advanced features into the convolutional layers of 1×1 and 3×3 . XH_{input} and XL_{input} have the same number of channels, $V = [V_1, V_2, \dots, V_c]$ is the set of convolutional kernels, where V_c is the parameter of the c -th convolution kernel. We were able to calculate output $U = [u_1, u_2, \dots, u_c]$:

$$u_c = v_c * X_{input} = \sum_{i=1}^c (v_c^i) * x^i \quad (3)$$

where $v_c = [v_c^1, v_c^2, \dots, v_c^c]$ and $X_{input} = [x^1, x^2, \dots, x^c]$, $X_{input} \in (XH_{input} \text{ or } XL_{input})$. The $*$ here represents convolution.

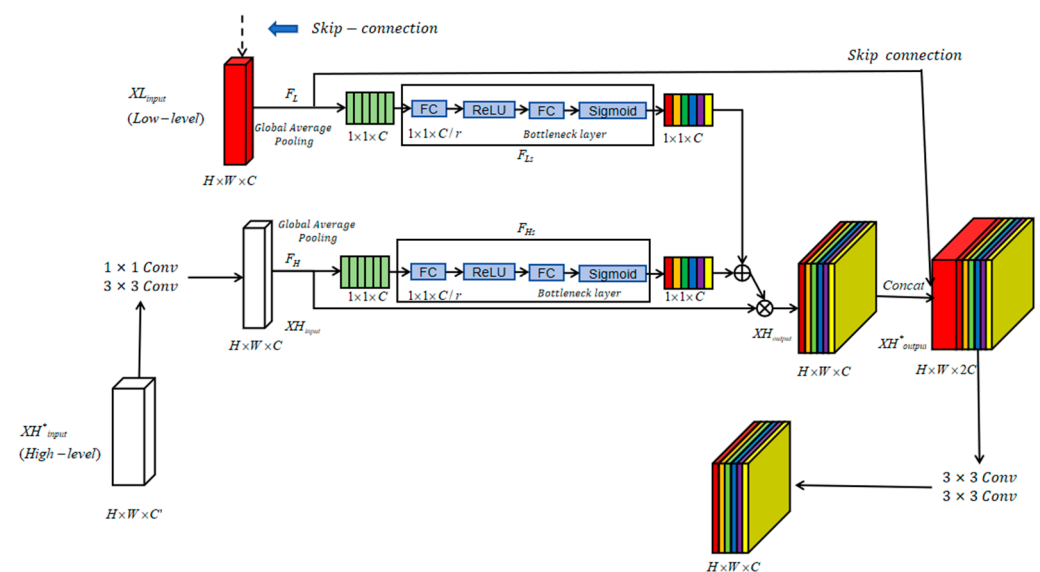


Figure 4. The structure of MFAB. MFAB uses two SE-Blocks to capture low-level and high-level feature maps, respectively. The final channel attention feature map is obtained through Concat connections. FC represents the fully connected layer. ReLU represents the Rectified Linear Unit.

Then, the global average pooling is utilized to compress each feature into a numerical column of $1 \times 1 \times C$, and channel level statistical information is generated. Formally, the statistical S_{c1} and S_{c2} are obtained by reducing the feature maps XH_{input} and XL_{input} . The calculation method for the c -th element of S_1 and S_1 is:

$$S_{c1} = F_L(XL_{input}) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W u_c(i, j) \quad (4)$$

$$S_{c2} = F_H(XH_{input}) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W u_c(i, j) \quad (5)$$

H and W represent height and width, respectively, and u_c represents the feature maps of each channel. Then, a bottleneck layer with two fully connected (FC) layers and an activation function is used to reduce the complexity of the model and capture channel dependencies z_1 and z_2 .

$$z_1 = F_{Ls}(S_1, P) = \delta_1(P_1 \delta_2(P_2, S_1)) \quad (6)$$

$$z_2 = F_{Hs}(S_2, P) = \delta_1(P_1 \delta_2(P_2, S_2)) \quad (7)$$

here, P_1 and P_2 represent fully connected layers, and δ_1 and δ_2 represents the sigmoid function and ReLU activation function, respectively. Then, we combine the channel outputs of low-level and high-level features using functions.

$$z = F_{add}(\cdot) = z_1 + z_2 \quad (8)$$

XH_{output} is obtained by rescaling T with activated V .

$$\tilde{X}H_{output-c} = F_{scale}(T_c, V_c) = V_c T_c \quad (9)$$

The final output X_{output} of MFAB is obtained by two 3×3 convolutional layers that capture semantic information.

3.5. Dataset Production

The dataset used in this experiment was produced by Northwest A&F University. To enhance dataset diversity, data were collected at the Baishui, Luochuan, and Qingcheng Apple Experiment Stations of Northwest A&F University. The dataset was primarily obtained under sunny and well-lit conditions, with some images collected on rainy or cloudy days. The diverse collection conditions further increased the diversity of the dataset. The dataset includes five common apple leaf pathological image data: 411 images of alternaria blotch, 435 images of brown spot, 370 images of grey spot, 375 images of mosaic, and 438 images of rust, totaling over 2029 images.

After annotation was completed, 1623 images were randomly selected as the training set, and images were enhanced using the algorithm's data augmentation library, a Python library used for image enhancement. The purpose of image enhancement is to create new data samples from existing datasets. Increasing training samples can help the model learn more effective features and effectively prevent overfitting. A total of 12 enhancement methods were used to expand the dataset, including horizontal flipping, vertical flipping, random flipping (maximum 90 degrees), random brightness and contrast adjustment, contrast limited adaptive histogram equalization, random cropping, resizing, random RGB value changes, blurring, Gaussian noise, tone separation, and random adjustment of image tone and saturation values. The expanded training set had a total of 21,099 images, and the expanded effect is shown in Figure 5.

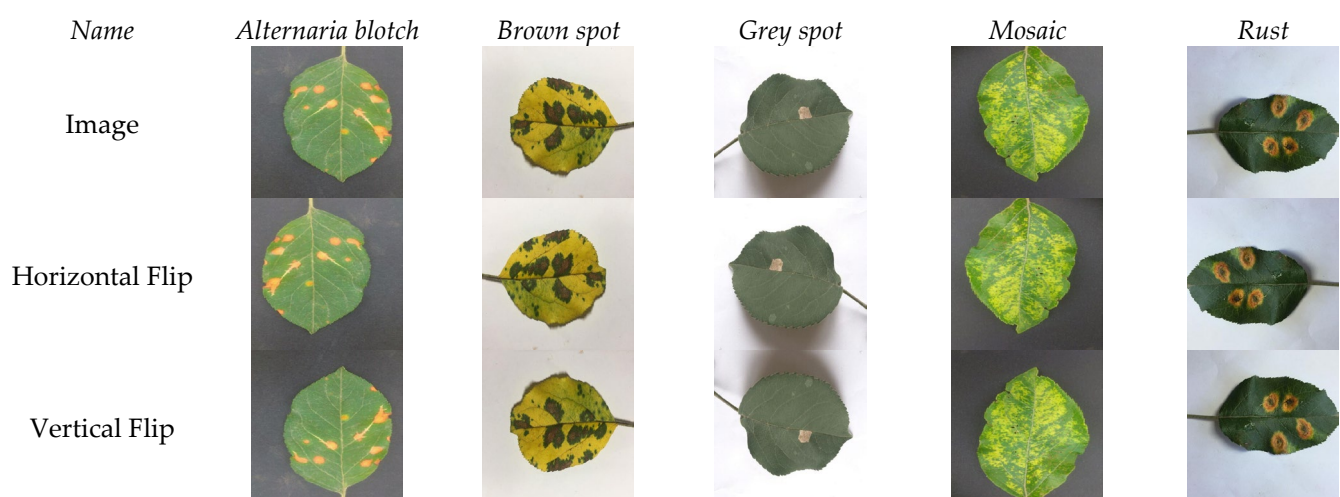


Figure 5. Cont.

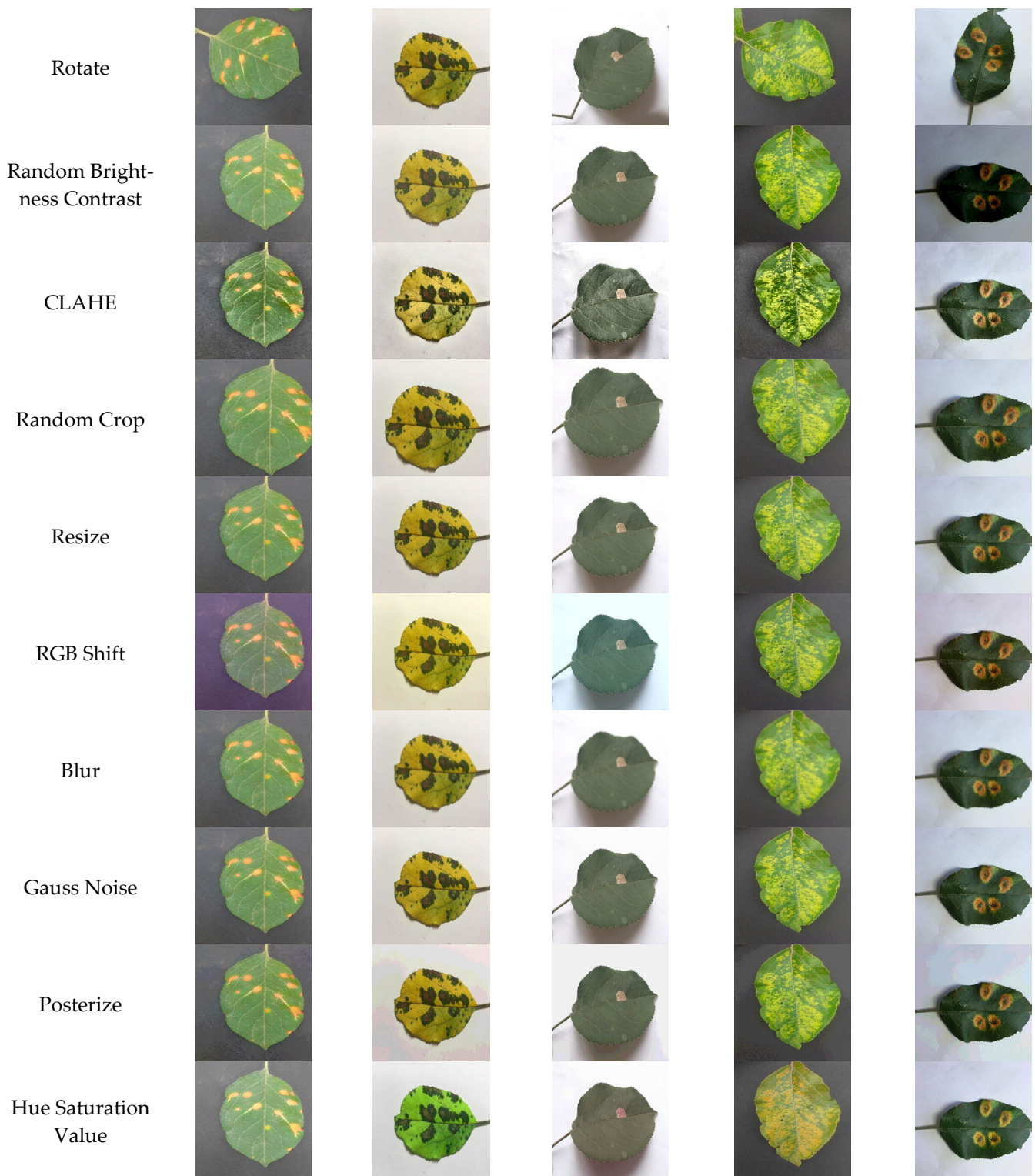


Figure 5. Disease image.

3.6. Metrics

accuracy: the proportion of correctly predicted quantity, in positive and negative examples, to the total quantity can be expressed by the formula:

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (10)$$

precision: based on the predicted results, the proportion of correctly predicted samples with positive examples is:

$$precision = \frac{TP}{TP + FP} \quad (11)$$

recall: based on actual samples, the proportion of correctly predicted positive samples to the total of actual positive samples.

$$recall = \frac{TP}{TP + FN} \quad (12)$$

F-score is a type of evaluation metric that harmonizes precision and recall:

$$F\text{-score} = (1 + \alpha^2) \frac{P \times R}{\alpha^2 \times P + R} \quad (13)$$

when $\alpha = 1$ is an F1 score, which is the harmonic mean of precision and recall (both are equally important). When $\alpha = 0.5$, it is an F0.5-score (precision is more important than recall). When $\alpha = 2$, it is an F2-score (recall is more important than precision).

For classification problems, examples can be divided into four scenarios based on the combination of their true category and the learner's predicted category: True Positive, False Positive, True Negative, and False Negative. *TP*, *FP*, *TN*, and *FN* represent the corresponding number of examples, respectively. The confusion matrix of the classification results is shown in Table 1.

Table 1. Basic concepts of positive and negative.

	Actual Positive	Actual Negative
Predictive positive	(True Positive) <i>TP</i>	(False Positive) <i>FP</i>
Predictive negative	(False Negative) <i>FN</i>	(True Negative) <i>TN</i>

Intersection over Union (*IoU*) is a standard for evaluating the accuracy of detecting corresponding objects in a specific dataset. Any task that obtains a bounding box in the output can be measured using *IoU*.

$$IoU = \frac{|A \cap B|}{|A \cup B|} \quad (14)$$

This article uses *Dice Loss* as the final loss function.

Dice Loss is named after the *Dice* coefficient, a metric used to assess the similarity between two samples. The larger the value, the more similar the samples are. The mathematical formula for the *Dice* coefficient is as follows:

$$Dice = \frac{2|A \cap B|}{|A| + |B|} \quad (15)$$

Dice Loss can be calculated as follows:

$$DiceLoss = 1 - Dice = 1 - \frac{2|A \cap B|}{|A| + |B|} \quad (16)$$

A and *B* represent the predicted binary image and the ground truth binary image, respectively.

3.7. Experimental Setting

The operating system of the experimental computer is Linux 5.15.0-56-generic, kernel version # 63-20.04.1-Ubuntu SMP Tue Nov 22 21:24:20 UTC 2022, hardware platform x86_64. Build experimental modules in Python 3.8.5. CUDA version is 11.7. Training batch_size is 32, num_workers are 8. The threshold of the evaluation indicator is 0.5. The initial learning rate is 0.0001, which is adjusted to 1×10^{-5} after 25 training sessions.

The evaluation indicators used in the experiment are Dice Loss, mIoU, accuracy, precision, recall, and F-score. Each experiment is trained 100 times. This paper first uses resnet101 as the encoder. Under the same encoder, U-Net, U-Net++, MA-Net, Link-Net, FPN, PSP-Net, PAN, DeepLabV3, and DeepLabV3+ are selected as the main structures. After experimental comparison, it is proven that the segmentation effect of MA-Net is superior. Afterwards, using MA-Net as the main structure, dpn98, se_resnet101, resnext101_32x8d, timm-regnetx_006, timm-regnety_006, densenet169, timm-gernet_m, mit_b2, and mit_b3 were used as encoders, respectively. After experimental comparison, it was proven that the segmentation effect is superior when using mit_b2 as the encoder. After selecting mit_b2 as the encoder, we replaced the original attention module PAB of MA-Net with attention modules SE, ECA, and CBAM. After experimental comparison, it was proven that the segmentation effect is superior when using ECA. Finally, we used cross validation to further evaluate the performance of the improved model.

4. Results

The feature extraction network model selected for the experiment is the resnet101 encoder, with different networks selected for the main structure and 100 training iterations. U-Net, a popular model in medical image segmentation, had an mIoU of 0.97960 and a Dice Loss of 0.01055. The mIoU of U-Net++ was 0.97875 and the Dice Loss was 0.01103, indicating that the redesigned skip connections and multi-depth U-Net collaborative learning do not have significant advantages in leaf lesion segmentation and recognition. The results of Link-Net and DeepLabV3 were similar, with mIoU values of 0.97841 and 0.97839, respectively. The results of PAN and DeepLabV3+ were worse, with mIoU values of 0.97514 and 0.97549, respectively. The results of FPN and PSP-Net were the worst, with mIoU values of 0.94500 and 0.94467, respectively. After comparison, it can be concluded that the result with MA-Net as the main structure was the best, with an mIoU of 0.97973 and a Dice Loss of 0.01048. This could be because MA-Net introduced PAB and MFAB. Our module has further improved MA-Net by using MiT as the encoder and replacing PAB with ECA, resulting in better results, with an mIoU of 0.97973 and a Dice Loss of 0.00947, as shown in Table 2.

Table 2. Comparison of Different Network Experiments Based on Resnet101 Encoder.

Method	Dice Loss	mIoU	Accuracy	Precision	Recall	Fscore
U-Net	0.01055	0.97960	0.99648	0.98945	0.98945	0.98945
U-Net++	0.01103	0.97875	0.99633	0.98898	0.98898	0.98898
Link-Net	0.01121	0.97841	0.99626	0.98880	0.98879	0.98879
FPN	0.02922	0.94500	0.99027	0.97080	0.97080	0.97080
PSP-Net	0.02943	0.94467	0.99021	0.97064	0.97064	0.97064
PAN	0.01297	0.97514	0.99571	0.98713	0.98713	0.98713
DeepLabV3	0.01122	0.97839	0.99627	0.98881	0.98881	0.98881
DeepLabV3+	0.01274	0.97549	0.99576	0.98728	0.98727	0.98728
MA-Net	0.01048	0.97973	0.99651	0.98952	0.98952	0.98952
Ours	0.00947	0.98159	0.99684	0.99053	0.99053	0.99053

We continued to use MA-Net as the structure, selecting different feature extraction networks as encoders. The experimental results of dpn98, se_resnet101, and timm-gernet_m were not good, with mIoU values of 0.97614, 0.97642, and 0.97574, respectively. The experimental results of resnext101_32x8d, timm-regnetx_006, and timm-regnety_006 were better, but did not exceed resnet101, with mIoU values of 0.97836, 0.97915, and 0.97705, respectively. The experimental results of densenet169, mit_b3, and mit_b2 all exceeded resnet101, with mIoU values of 0.98025, 0.98098, and 0.98123, respectively. After comparison, it can be concluded that using mit_b2 as the encoder produced the highest mIoU, as shown in Table 3.

Table 3. Comparison of Experiments on Different Encoders Based on MA-Net.

Encoder	Dice Loss	mIoU	Accuracy	Precision	Recall	Fscore
resnet101	0.01048	0.97973	0.99651	0.98952	0.98952	0.98952
dpn98	0.01229	0.97614	0.99591	0.98772	0.98772	0.98772
se_resnet101	0.01220	0.97642	0.99594	0.98781	0.98781	0.98781
resnext101_32x8d	0.01118	0.97836	0.99628	0.98883	0.98882	0.98883
timm-regnetx_006	0.01078	0.97915	0.99641	0.98923	0.98923	0.98922
timm-regnety_006	0.01186	0.97705	0.99605	0.98814	0.98814	0.98814
densenet169	0.01019	0.98025	0.99661	0.98982	0.98982	0.98982
timm-gernet_m	0.01251	0.97574	0.99583	0.98750	0.98750	0.98750
mit_b3	0.00982	0.98098	0.99673	0.99019	0.99018	0.99018
mit_b2	0.00969	0.98123	0.99677	0.99031	0.99031	0.99031

Next, we used MA-Net as the structure and mit_b2 as the encoder, and replaced the PAB module of MA-Net. The structure of SENet is composed of SEblock blocks, and the core idea is to adaptively adjust the weights of each channel by learning the relationships between channels, thereby improving the network's expressive power. The CBAM module has two sub-modules: the Channel Attention Module and the Spatial Attention Module, which can perform attention operations in both spatial and channel dimensions to improve the model's attention to important features. The ECA module extracts important information from the feature map by weighting the channel dimensions of the input feature map with attention. Compared to the original version, the experimental results of all three attention modules improved to a certain extent. After comparison, it can be concluded that the mIoU generated while using ECANet as the attention mechanism was the highest, as shown in Table 4.

Table 4. Comparative experiment of different attention modules based on MA-Net and mit_b2.

Attention	Dice Loss	mIoU	Accuracy	Precision	Recall	Fscore
PAB	0.00969	0.98123	0.99677	0.99031	0.99031	0.99031
SENet	0.00964	0.98131	0.99679	0.99037	0.99037	0.99037
CBAM	0.00968	0.98124	0.99677	0.99032	0.99032	0.99032
ECANet	0.00947	0.98159	0.99684	0.99053	0.99053	0.99053

Next, we used different network structures as the main structure and resnet101 as the encoder. FPN and PSP-Net were almost unable to accurately segment lesions, simply because a large amount of background raises the mIoU. U-Net, U-Net++, and LINK-Net performed well in segmenting five types of lesions, but were prone to misidentification of grey spot, often misidentifying the withered areas of leaves as grey spot. PAN and DeepLabV3+ both tended to misidentify grey spot as alternaria blotch. DeepLabV3 and MA-Net had better segmentation and recognition of leaf lesions. After comparison, it can be seen that when MA-Net is the main structure, the image segmentation effect for the five types of lesions was the best. Our improved model for MA-Net performed better, as shown in Figure 6.

Next, we used MA-Net as the main structure and different feature extraction networks as encoders for comparative experiments. Dpn98 and timm-gernet_m could not distinguish between alternaria blotch and grey spot. Se_resnet101 and timm-regnety_006 had poor segmentation of grey spot and were prone to misidentifying it as alternaria blotch. The segmentation effect of timm-regnetx_006 on mosaic disease was average. Resnext101_32x8d, timm-regnetx_006, and densenet169 showed a tendency to misidentify the withered parts of leaves as grey spot. The segmentation results of mit_b2 and mit_b3 were better, but mit_b3 sometimes misidentified brown spot as grey spot. From the resulting image, it can be seen that using mit_b2 as the encoder had the best segmentation effect on the five types of lesions, as shown in Figure 7.

Using MA-Net as the main structure and mit_b2 as the encoder, experiments were conducted using different attention modules. SENet, CBAM, and ECA showed good segmentation and recognition effects on the five types of lesions. However, SENet misidentified alternaria blotch and brown spot, while CBAM misidentified brown spot and grey spot. From the resulting images, it can be seen that using ECA as the attention module had the best segmentation effect on the five types of lesions, as shown in Figure 8.

In deep learning, performance evaluation of models is crucial. However, the traditional single model evaluation method is often affected by the data set partition, and cannot give an accurate performance evaluation. To address this issue, cross-validation has been widely adopted as a more accurate and reliable evaluation method. Cross-validation provides a more accurate estimate of the model's generalization ability by dividing the original dataset into multiple subsets and using these subsets for multiple model training and performance evaluations. This article uses the threefold cross-validation method, which divides the original dataset into three subsets. Each time, two subsets are selected as the training set, and the remaining subset is used as the validation set. This process is repeated three times, each time selecting a different subset as the validation set. The results are shown in Table 5.

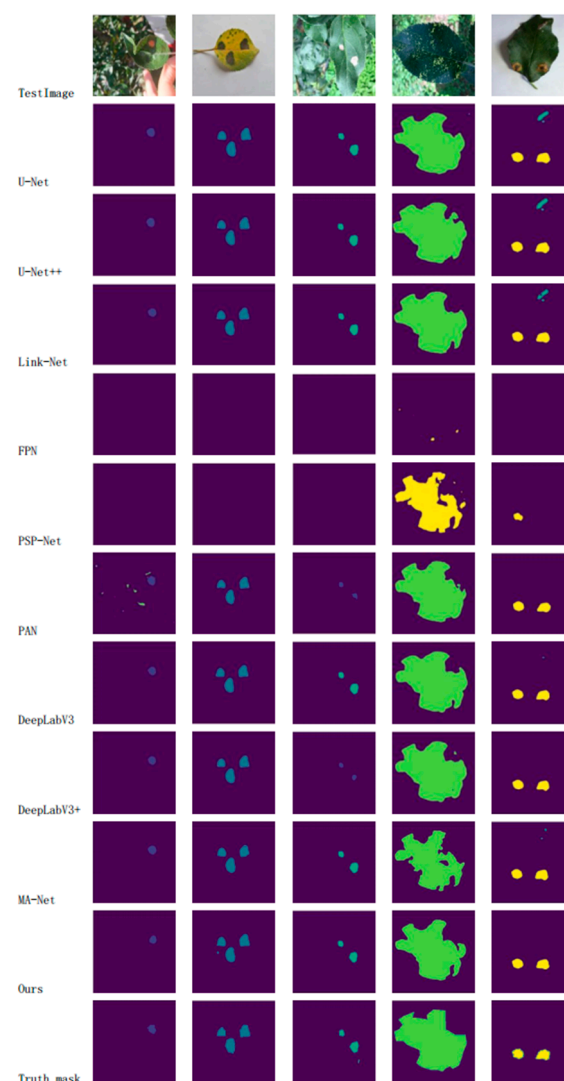


Figure 6. Example of the results of lesion segmentation using different methods on the test dataset. Purple represents alternaria blotch, blue represents brown spot, cyan represents grey spot, green represents mosaic, and yellow represents rust.

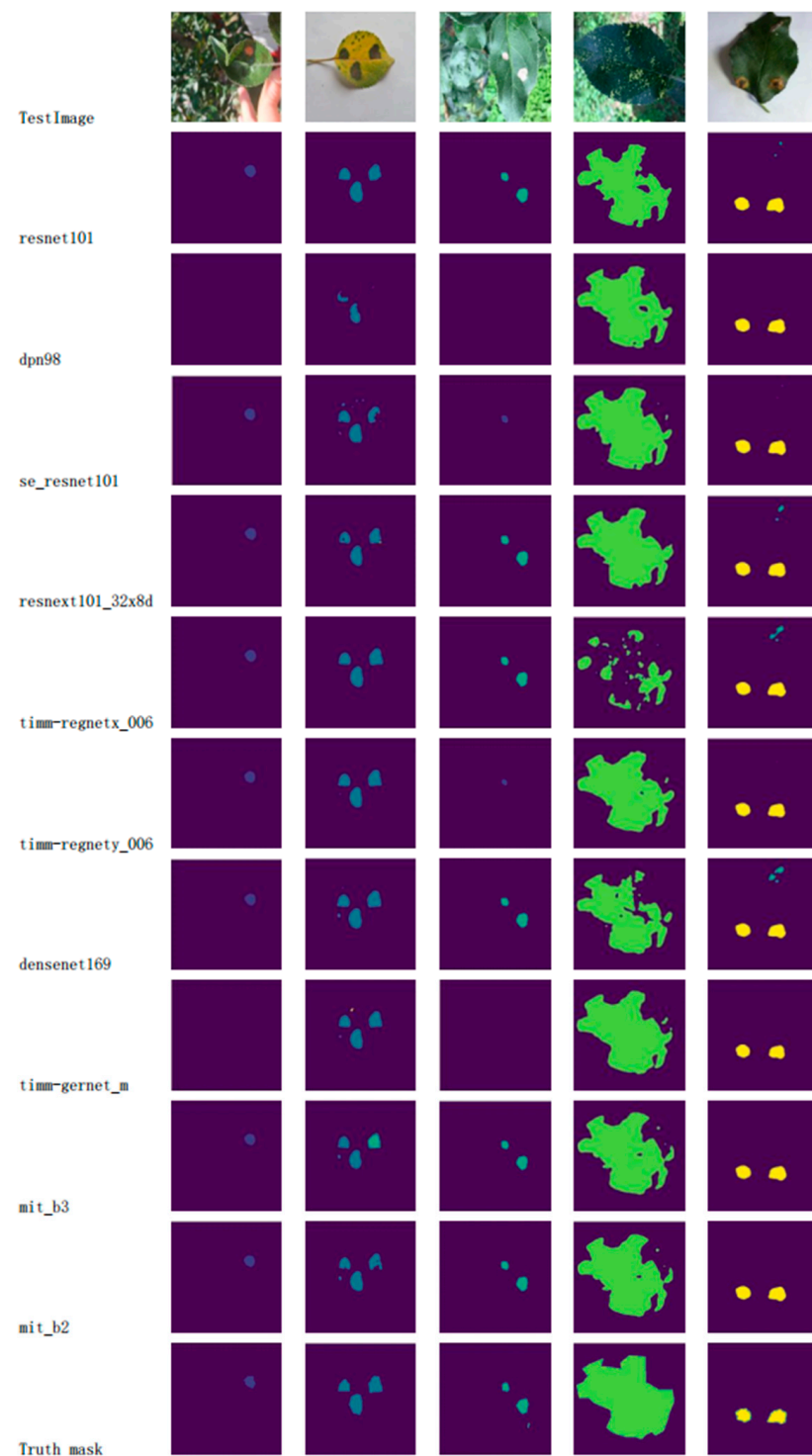


Figure 7. Example of the results of lesion segmentation using different encoders on the test dataset. Purple represents alternaria blotch, blue represents brown spot, cyan represents grey spot, green represents mosaic, and yellow represents rust.

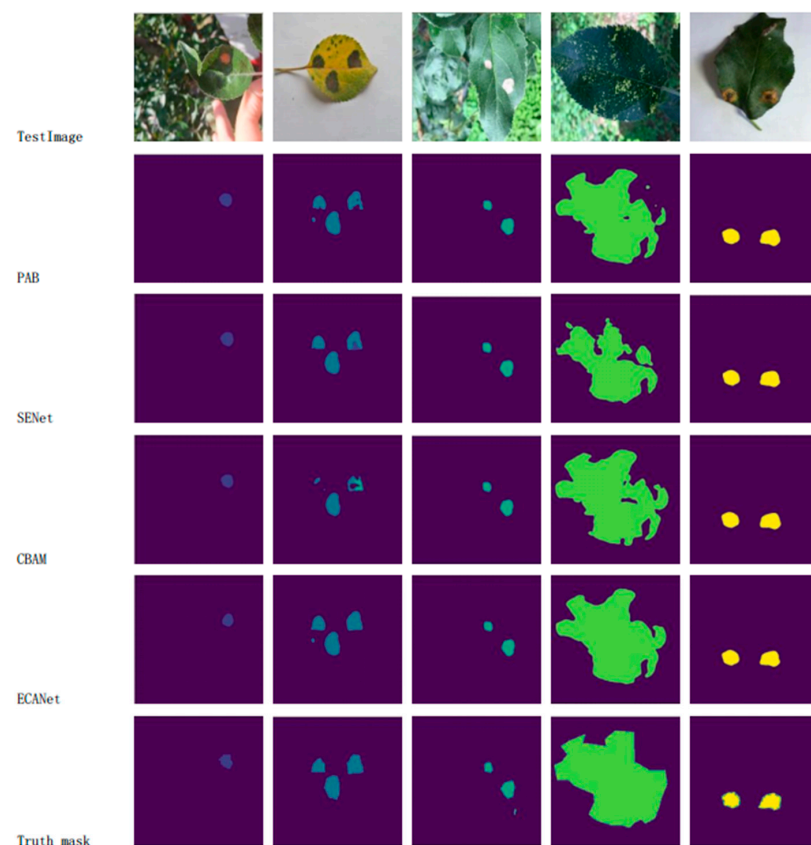


Figure 8. Example of the results of lesion segmentation on different attention modules on the test dataset. Purple represents alternaria blotch, blue represents brown spot, cyan represents grey spot, green represents mosaic, and yellow represents rust.

Table 5. Result of 3-fold cross-validation.

Fold	Dice Loss	mIoU	Accuracy	Precision	Recall	F-Score
Fold1	0.00988	0.98001	0.99653	0.99010	0.99010	0.99010
Fold2	0.01083	0.97932	0.99639	0.98917	0.98917	0.98917
Fold3	0.01224	0.97633	0.99592	0.98777	0.98777	0.98777

To evaluate and optimize the performance of machine learning models, a performance graph is used to visually observe the performance of the model at different training stages. Figure 9 shows the performance of the model on the validation set. It is obvious that the model has undergone good training and achieved stability.

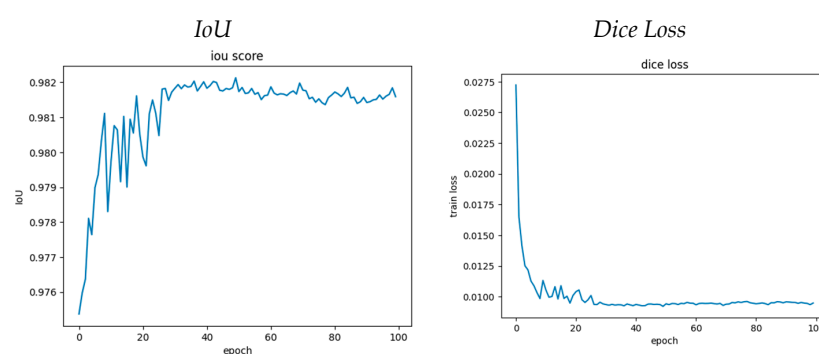


Figure 9. Performance graphs.

To evaluate the generalization performance of a given algorithm trained on a specific dataset, the dataset was subjected to threefold cross-validation.

5. Discussion

The automatic segmentation of plant lesions helps agricultural experts in clinical diagnosis to accurately outline the lesions and assist farmers in immediate treatment of crops. This article designs a plant leaf lesion segmentation network based on an improved MA-Net. We introduced a Mix Vision Transformer and ECA into this method. To demonstrate the superiority of this method, this paper first used resnet101 as the encoder. Under the same encoder, U-Net, U-Net++, MA-Net, Link-Net, FPN, PSP-Net, PAN, DeepLabV3, and DeepLabV3+ were selected as the main structures. After experimental comparison, it was proven that the segmentation effect of MA-Net is superior, as the mIoU reached 0.97973. Afterwards, using MA-Net as the main structure, dpn98, se_resnet101, resnext101_32x8d, timm-regnetx_006, timm-regnety_006, densenet169, timm-gernet_m, mit_b2, and mit_b3 were used as encoders, respectively. After experimental comparison, it was proven that the segmentation effect is superior when using mit_b2 as the encoder, as the mIoU reached 0.98123. After selecting mit_b2 as the encoder, we replaced the original attention module PAB of MA-Net with attention modules SE, ECA, and CBAM. After experimental comparison, it was proven that the segmentation effect is superior when using ECA, as the mIoU reached 0.98159. Compared to Res-block, MiT consumes less time during the training process, can better capture global and contextual information in images, and has better robustness and scalability. Compared to PAB, which has higher computational complexity and requires a large amount of computing resources, ECA reduces the number of model parameters and computation by learning the correlation between channels. ECA also has better denoising ability.

ECA learns channel attention through a 1D convolutional layer, thereby reducing computational complexity. PAB usually contains two fully connected layers, so the computational complexity of PAB is relatively high. Due to the low computational complexity of ECA, the required time is also relatively small. ECA itself does not add additional parameters, and the number of parameters mainly depends on the number of channels in the input feature map. The number of parameters in ECA is relatively small, with only one convolutional layer parameter. PAB has a large number of parameters, including the parameters of two fully connected layers.

Parameters are the variables that are learned and optimized by the model during the training process. They are specific to the model and are adjusted based on the available data. Parameters help the model to capture patterns and trends in the data, and they are typically initialized with random values. Parameters are typically optimized using algorithms like gradient descent, which update the values of the parameters to minimize the error on the training set. The number of parameters is related to the number of channels in the input feature map, assuming there are C channels in the input feature map. As shown in Table 6, ECA has certain advantages in computational complexity, computation time, and number of parameters.

Table 6. Comparison of attention modules. The size of the input feature map is $H \times W$, C is the number of channels, D is dimension, r is the compression ratio (usually 16), the size of the convolution kernel is $k \times k$, the number of attention heads is h , and the sequence length is n .

Attention	Computational Complexity	Time Taken	Parameters
PAB	$O(D^2 + D^2 + n^2 \times h + n \times D \times h)$	0.01294 s	$5C^2 + C^3/8$
SENet	$O(H \times W \times C + C^2)$	0.00032 s	$2 \times C/r$
CBAM	$O(C \times k1 \times k1 \times H \times W + H \times W \times k2 \times k2 \times C)$	0.00038 s	$C \times (C + 2) + k^2 \times C$
ECANet	$O(C \times D + D^2)$	0.00026 s	$2 \times C$

6. Conclusions

This article studies the image segmentation problem of plant leaf lesions. This article proposes an improved MA-Net network architecture for plant disease spot segmentation. We annotated and expanded the apple leaf image dataset captured by Northwest A&F University in 2019, replaced the encoder and attention module of traditional MA-Net, and conducted comparative experiments. The experimental results show that the algorithm proposed in this article can achieve good segmentation results when processing the 2019 apple leaf dataset from Northwest A&F University. Compared with traditional image processing methods, the algorithm proposed in this paper has advantages in accuracy and robustness. This method helps farmers in their crop control process. However, there are also shortcomings in the experiment. In complex outdoor environments, the performance of segmentation algorithms is affected by issues such as occlusion, uneven lighting, and shadows. Different growth stages and varieties may exhibit different disease characteristics, posing challenges for identification. Deep learning technology requires a large amount of annotated data, and in certain regions or specific crops there are many types of diseases and a large amount of annotation work. Future research might focus on the use of deep learning techniques to integrate multiple modalities of data (such as spectra, textures, shapes, etc.) to improve the accuracy and robustness of disease recognition. In the absence of a large amount of annotated data, unsupervised and semi-supervised learning methods such as self-supervised learning, transfer learning, and weakly-supervised learning might be further researched to reduce dependence on a large amount of annotated data. In order to achieve real-time diagnosis on devices with limited resources, research might be conducted on how to prune deep learning models to reduce model size and computational complexity, making them more lightweight.

Author Contributions: Conceptualization, T.Y.; methodology, T.Y.; validation, T.Y.; formal analysis, T.Y.; writing—original draft preparation, T.Y.; writing—review and editing, J.L. and Y.W.; visualization, T.Y.; supervision, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in <https://aistudio.baidu.com/datasetdetail/259166> (accessed on 10 February 2024).

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

References

1. Xiaolei, B.; Yuehua, W.; Hanbing, Z. *Development of Agriculture and Forestry and Food Safety*; Jilin People's Publishing House: Jilin, China, 2017; p. 123.
2. Xu, L.; Zhao, Y.; Liu, P.; Ban, X.; Zhao, J.; Dai, F. Application of an improved lightweight YOLOv5 model to wheat stripe rust detection. In Proceedings of the 2023 11th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Wuhan, China, 25–28 July 2023.
3. Tian, Y.; Li, T.; Li, C.; Paio, Z.; Sun, G.; Wang, B. Grape disease image recognition method based on support vector machine. *J. Agric. Eng.* **2007**, *23*, 175–180.
4. Zongru, L.; Dongjian, H. Research on Apple Disease Identification Technology Based on Mobile Phone Image Analysis. *Comput. Eng. Des.* **2010**, *31*, 3051–3053 + 3095.
5. Kaur, R.; Kang, S.S. An enhancement in classifier support vector machine to improve plant disease detection. In Proceedings of the 2015 IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE), Amritsar, India, 1–2 October 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 135–140.
6. Shuangping, H.; Long, Q.; Xu, M.; Kunnan, X.; Wenjuan, W. Classification method of rice blast disease severity based on hyperspectral imaging. *J. Agric. Eng.* **2015**, *31*, 212–219.
7. Padol, P.B.; Yadav, A.A. SVM classifier based grape leaf disease detection. In Proceedings of the 2016 Conference on Advances in Signal Processing (CASP), Pune, India, 26–30 June 2016; pp. 175–179.
8. Ma, Y.; Feng, Q.; Yang, M.; Li, M. Detection of wine grape leaves based on HOG. *Comput. Eng. Appl.* **2016**, *52*, 158–161.

9. Islam, M.; Dinh, A.; Wahid, K.; Bhowmik, P. Detection of potato diseases using image segmentation and multiclass support vector machine. In Proceedings of the 2017 IEEE 30th Canadian conference on electrical and computer engineering (CCECE), Windsor, ON, Canada, 30 April–3 May 2017; pp. 1–4.
10. Zhen, W.; Guikun, C.; Jinxing, W.; Xincheng, H.; Furui, G.; Xinhua, D. Detection of IK SVM rice blast spores based on HOG characteristics. *J. Agric. Mach.* **2018**, *49*, 387–392.
11. Sun, Y.; Jiang, Z.; Zhang, L.; Dong, W.; Rao, Y. SLIC_SVM based leaf diseases saliency map extraction of tea plant. *Comput. Electron. Agric.* **2019**, *157*, 102–109. [[CrossRef](#)]
12. Ruijin, Z. Research on identification of barley diseases based on support vector machines. *Heilongjiang Sci.* **2019**, *10*, 1–3.
13. Hinton, G.E.; Salakhutdinov, R.R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313*, 504–507. [[CrossRef](#)] [[PubMed](#)]
14. Sladojevic, S.; Arsenovic, M.; Anderla, A.; Culibrk, D.; Stefanovic, D. Deep neural networks based recognition of plant diseases by leaf image classification. *Comput. Intell. Neurosci.* **2016**, *2016*, 3289801. [[CrossRef](#)] [[PubMed](#)]
15. Brahimi, M.; Boukhalfa, K.; Moussaoui, A. Deep learning for tomato diseases: Classification and symptoms visualization. *Appl. Artif. Intell.* **2017**, *31*, 299–315. [[CrossRef](#)]
16. Xiao, Z.; Liu, H. Adaptive feature fusion and fast recognition of typical potato disease images. *J. Agric. Mach.* **2017**, *48*, 26–32.
17. Veeraballi, R.K.; Nagugari, M.S.; Annavarapu, C.S.R.; Gownipuram, E.V. Deep Learning Based Approach for Classification and Detection of Papaya Leaf Diseases. In Proceedings of the International Conference on Intelligent Systems Design and Applications, Vellore, India, 6–8 December 2018; Springer: Cham, Switzerland, 2018; pp. 291–302.
18. Zhang, X.; Qiao, Y.; Meng, F.; Fan, C.; Zhang, M. Identification of maize leaf diseases using improved deep convolutional neural networks. *IEEE Access* **2018**, *6*, 30370–30377. [[CrossRef](#)]
19. Silva Abade, A.; Almeida, A.P.G.S.; Barros Vidal, F. Plant Diseases Recognition from Digital Images using Multichannel Convolutional Neural Networks. In Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (5: VISAPP), Prague, Czech Republic, 25–27 February 2019; pp. 450–458.
20. Hu, Z.; Yang, H.; Huang, J.; Xie, Q. Fine grained tomato disease recognition based on attention residual mechanism. *J. South China Agric. Univ.* **2019**, *40*, 124–132.
21. Sharma, P. Performance analysis of deep learning CNN models for disease detection in plants using image segmentation. *Inf. Process. Agric.* **2019**, *7*, 566–574. [[CrossRef](#)]
22. Jiang, F.; Lu, Y.; Chen, Y.; Cai, D.; Li, G. Image recognition of four rice leaf diseases based on deep learning and support vector machine. *Comput. Electron. Agric.* **2020**, *179*, 105824. [[CrossRef](#)]
23. Jinxiu, H.; Ran, L.; Hongxia, D.; Haifang, L. Leaf Disease Identification Based on Fusion Channel Information Attention Network. *Comput. Eng. Appl.* **2020**, *56*, 124–129.
24. Su, D.; Qiao, Y.; Kong, H.; Sukkarieh, S. Real time detection of inter-row ryegrass in wheat farms using deep learning. *Biosyst. Eng.* **2021**, *204*, 198–211.
25. Yongbo, L.; Liang, H.; Yan, C.; Jiangyun, T.; Bo, L. A segmentation algorithm for corn leaf lesions based on U-Net. *Chin. J. Agron.* **2021**, *37*, 88–95.
26. Changyuan, Z.; Zelin, H.; Miao, L.; Hualong, L.; Xuanjiang, Y.; Fei, L. Real time semantic segmentation model for crop disease leaves based on group attention module. *J. Agric. Eng.* **2021**, *37*, 208–215.
27. Su, D.; Kong, H.; Qiao, Y.; Sukkarieh, S. Data augmentation for deep learning based semantic segmentation and crop-weed classification in agricultural robotics. *Comput. Electron. Agric.* **2021**, *190*, 106418.
28. Jiang, Z.; Dong, Z.; Jiang, W.; Yang, Y. Recognition of rice leaf diseases and wheat leaf diseases based on multi-task deep transfer learning. *Comput. Electron. Agric.* **2021**, *186*, 106184. [[CrossRef](#)]
29. Xiangpeng, F.; Jianping, Z.; Yan, X.; Xuan, P. Identification of corn diseases in complex backgrounds based on improved convolutional neural networks. *J. Agric. Mach.* **2021**, *37*, 210–217.
30. Shifang, S.; Yan, Q.; Grape, R.Y. Leaf disease identification and mobile application based on transfer learning. *J. Agric. Eng.* **2021**, *37*, 127–134.
31. Li, Q.; Nan, M.; Xin, Z.; Xueying, Y.; Shoucheng, W.; Jiyong, G.; Zhiqiang, W. Corn disease image recognition based on attention mechanism asymmetric residual network and transfer learning. *Sci. Technol. Eng.* **2021**, *21*, 6249–6256.
32. Ning, Z.; Huarui, W.; Xiao, H.; Yisheng, M. A tomato disease recognition method based on multi-scale and attention mechanism. *Zhejiang Agric. J.* **2021**, *33*, 1329–1338.
33. Junjie, W.; Lijun, Q.; Zhongao, L.; Jiarui, Z.; Hao, Z. GoogLeNet Orchard Disease and Pest Identification and Grading Based on Transfer Learning. *J. China Agric. Univ.* **2021**, *26*, 209–221.
34. Wenxia, B.; Gang, W.; Gensheng, H.; Dongyan, Z.; Linsheng, H. Identification of Apple Leaf Diseases Based on Improved Convolutional Neural Networks. *J. Anhui Univ. (Nat. Sci. Ed.)* **2021**, *45*, 53–59.

35. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *arXiv* **2021**, arXiv:2105.15203. [[CrossRef](#)]
36. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, PR, USA, 17–19 June 1997; pp. 11531–11539.

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.