# Lightweight Grape Leaf Disease Recognition Method Based on Transformer Framework

**Enxu Zhang**
15636090983@163.com

Universities of Shaanxi Province, Xijing University

**Guowei Qi**
Universities of Shaanxi Province, Xijing University

**Fei Li**
Universities of Shaanxi Province, Xijing University

**Cheng Lv**
Universities of Shaanxi Province, Xijing University

**Ning Zhang**
Universities of Shaanxi Province, Xijing University

**Article**

**Keywords:** Grape Leaf Disease, Deep Learning, Lightweight Transformer Model, Attention Mechanisms

**Posted Date:** September 9th, 2024

**DOI:** https://doi.org/10.21203/rs.3.rs-4858639/v1



**Additional Declarations:** No competing interests reported.

**Version of Record:** A version of this preprint was published at Scientific Reports on August 7th, 2025.
See the published version at https://doi.org/10.1038/s41598-025-13689-7.

Article

# Lightweight Grape Leaf Disease Recognition Method Based on Transformer Framework

**Enxu Zhang [1], Guowei Qi [1] , Fei Li [1], Cheng Lv [1] and Ning Zhang [1,*]**

[1] Engineering Research Center of Hydrogen Energy Equipment& Safety Detection, Universities of Shaanxi Province, Xijing University, Xi'an, 710123, China.
* Correspondence

**Abstract:** Grape disease image recognition is a crucial part of agricultural disease detection, and accurate identification of grape leaves plays a vital role in agricultural production. This study proposes a deep learning-based method for grape disease classification and recognition to address issues such as the complexity of grape disease features, uneven distribution of disease features, and data imbalance. First, the adversarial generative network FastGAN is used to generate grape disease images to enrich the sample information for different categories in the dataset and address the data imbalance problem. Then, a novel Transformer structure called LVT Block and a CNN structure called MARI Block are proposed to process the global and local information of images, respectively. Dense connections between these structures result in the DLVT Block, which leads to the lightweight neural network model DLVTNet. Additionally, a lightweight self-attention mechanism combined with CNN called CLSHSA is introduced, which maintains high recognition performance while reducing the model size. Moreover, a multi-scale attention mechanism (MELA) is proposed, which combines positional and multi-scale information to obtain attention weights. Experimental results show that this method achieves an average recognition accuracy of 98.48% in grape leaf disease detection, outperforming mainstream CNN and Transformer models, and effectively focuses on disease areas in leaf images. The method also demonstrates high recognition accuracy in tomato disease detection, indicating good generalization ability and suitability for detecting and recognizing various leaf diseases. The proposed method provides an effective solution for detecting and recognizing grape and other plant leaf diseases, offering a new research approach that combines CNN and Transformer structures.

**Keywords:** Grape Leaf Disease, Deep Learning, Lightweight Transformer Model, Attention Mechanisms

## 1. Introduction

Grapes hold a significant position in global fruit production, being widely cultivated in the Northern Hemisphere. However, during grape cultivation, grapevines are susceptible to various diseases triggered by factors such as bacteria, viruses, and environmental conditions, severely impacting grape production [1]. Among the various diseases in grape cultivation, leaf diseases are the most common and have a significant impact on plant growth, development, yield, and quality. These diseases are more likely to occur under humid conditions, such as during rainy seasons. How-ever, due to the diversity of grape diseases, complex occurrence conditions, significant differences between similar diseases, and subtle differences between dissimilar diseases, disease prevention and control face considerable challenges [2]. Traditional manual detection methods are affected by these conditions, resulting in low identification efficiency and accuracy, making it difficult to effectively prevent and control diseases. Additionally, in the early stages of grape disease occurrence, certain characteristics different from healthy leaves often appear on the leaf surface, providing a basis for using machine vision technology to identify diseases. Today, with the continuous development of machine vision technology, it is possible to utilize the characteristic manifestations of grape leaves to timely identify diseases and implement prevention measures, thereby effectively reducing excessive spraying of chemicals and minimizing environ-mental impact [3].

In the early stages, many scholars mainly used traditional machine learning algorithms to identify diseases with complex feature representations. Zhang et al. proposed a GA-SVM model using genetic algorithms to automatically obtain penalty factors and kernel functions for corn disease classification, showing improvement over traditional SVM models [4]. Tan et al. proposed a citrus surface defect recognition method based on KF-2D-Renyi and ABC-SVM, achieving high-precision recognition and classification of eight types of citrus surface defects through dark channel prior defogging, optimized 2D Renyi entropy threshold segmentation, and ABC-SVM classifier [5].

However, recognition based on traditional machine learning often relies on human experience to set features such as color, shape, and texture, making it difficult to effectively identify leaf diseases with complex disease feature representations. In recent years, deep learning technology has developed rapidly, and its powerful performance in feature ex-traction and recognition has gradually been discovered. Among them, CNN models can design feature extractors based on the training data by introducing local connectivity and weight sharing, achieving significant results in various diseases [6] . Chen et al. proposed a tomato leaf disease recognition framework based on ABCK-BWTR and B-ARNet, enhancing images through binary wavelet transform combined with Retinex denoising, optimizing KSW to separate leaves from the background, and using a du-al-channel residual attention network for recognition[7]. Goncharov et al. improved the deep Siamese network and single-layer perceptron classifier, expanding the database to five groups of grape, corn, and wheat images, effectively in-creasing the plant disease detection accuracy to 96% [8]. Bao et al. designed a lightweight SimpleNet model for automatic identification of wheat ear diseases in the natural scene images taken in the field, and connected the downsampled feature map output by the inverted residual block with the average pooling feature of the feature map of the input inverted residual block to achieve the fusion between features of different depths, reduce the loss of detailed features of wheat ear disease caused by the network during the downsampling process, solve the problem of feature extraction of the disappearance of disease features in the imaging process, and achieved an accuracy of 96.71% [9]. Atila et al. studied the use of the EfficientNet architecture for plant leaf disease classification on the PlantVillage dataset, achieving high classification accuracy on both original and enhanced datasets through transfer learning methods [10]. Lin proposed a lightweight CNN model called GrapeNet, effectively identifying grape disease symptoms at different stages using residual blocks and convolutional block attention modules, demonstrating superior classification performance compared to other classic models while significantly reducing parameter count and training time [11]. Fang et al. proposed a new network architecture called HCA-MFFNet, utilizing hard coordinated attention (HCA) allocated at different spatial scales and multi-feature fusion techniques to effectively identify corn leaf diseases in complex backgrounds, achieving a recognition accuracy of 97.75% [12]. Zhang et al. introduced the asymptotic non-local means algorithm (ANLM) to reduce im-age noise interference and proposed a multi-channel automatic-oriented recurrent attention network (M-AORANet) to address tomato leaf disease recognition issues to some ex-tent [13]. Sharma et al. proposed a deeper lightweight multi-class model DLMC-Net for plant leaf disease detection, introducing collective blocks and channel layers to avoid gradient vanishing problems, demonstrating better detection of multiple plant leaf diseases compared to other network models [14].

In network models based on deep learning, improving attention mechanisms can effectively enhance model recognition performance. Zhao et al. introduced the SE-Net attention mechanism into the ResNet-50 model to help extract effective channel information and combined it with a multi-scale feature extraction module to achieve tomato disease recognition [15]. Zhao et al. embedded the CBAM attention mechanism into the Inception network model to identify corn, potato, and tomato diseases [16]. Zeng et al. proposed a self-attention convolutional neural network (SACNN) and added it to the basic network to achieve crop plant disease recognition, obtaining good recognition results on the MK-D2 dataset [17]. Chen et al. proposed an attention module (LSAM) for use in MobileNet V2 to form the Mo-bile-Atten model, achieving an average accuracy of 98.48% for rice disease recognition in complex backgrounds [18].

In recent years, with the expansion of transformer applications, their powerful recognition capabilities in CNN models have provided new methods for agricultural disease recognition and detection. Vallabhajosyula et al. combined transformer and resnet9 models for leaf disease classification and detection, demonstrating superiority over main-stream CNN models [19]. Waheed et al. used vision transformer, Swin transformer, and Compact Convolutional Transformer (CCT) for tomato disease recognition and detection, achieving effective identification. Wei Li used the CBAM attention mechanism and Transformer framework add-ed to Resnet50 to achieve high-precision recognition of tomatoes, realizing effective classification on the PlantDoc dataset [20]. Chen et al. addressed the challenge of recognizing tomato leaf diseases by first using CyTrGan to generate a certain number of sample images to enrich the image dataset, then using a densely connected CNN network and Transformer structure, experimentally verifying effective recognition of tomato diseases [21]. Han et al. used a dual-channel Swin Transformer for image detection, with two channels receiving initial image input and edge information input respectively, demonstrating effective recognition of ligneous leaf diseases [22]. Hu et al. proposed a hybrid framework FOTCA to address the poor performance of trans-formers on small to medium-sized datasets, using adaptive Fourier neural operators and Transformer architecture to effectively extract global features and achieve effective disease recognition [23].

In the field of grape leaf disease recognition, there have also been numerous achievements. Ustad et al. developed an image processing-based grape plant disease detection and classification system using K-means clustering algorithm combined with SVM classifier [24]. Phookronghin et al. used self-organizing feature maps (SOFM) to extract diseased areas of grape leaves and used a 2-level simplified fuzzy ARTMAP (2L-SFAM) for grape leaf dis-ease recognition and classification[25]. Mohammed et al. established an artificial intelligence technique for grape leaf disease detection and classification, using K-means algorithm for segmentation, extracting texture features from segmented grape leaves, and then using multiple support vector machines and Bayesian classifiers to determine the type of grape leaf disease [26]. A, Miaomiao Ji et al. proposed a joint convolutional neural network architecture United-Model based on ensemble methods to achieve grape disease recognition, obtaining an average validation accuracy of 99.17%[27]. Suo et al. proposed a coordinated attention shuffle mechanism-asymmetric multi-scale fusion module network model (CASM-AMFMNet) by combining Gaussian filtering Sobel smoothing denoising Laplacian operator and CoAtNet network for effective recognition of grape leaf diseases [28]. Liu proposed a CNN model combining Inception structure and dense connection strategy, achieving effective recognition of grape diseases compared to GoogLeNet and ResNet-34 [29]. Cai et al. used binary wavelet transform combined with variable threshold method and NL-means improved MSR algorithm (VN-BWT) to enhance grape leaf images, and pro-posed a new Siamese network (Siamese DWOAM-DRNet) combining multi-branch residual module (DRM) and dual-factor weight optimization attention mechanism (DWOAM) for grape leaf disease classification, effectively improving disease recognition ability in complex backgrounds [30]. Karthik et al. proposed a novel dual-track feature fusion network called GrapeLeaf-Net, using InceptionResNet and CBAM for local feature extraction and Shuffle-Transformer for global feature extraction, achieving high recognition capability for grape leaf disease detection [31].

Although significant research achievements have been made in agricultural leaf disease recognition, some issues remain. In agricultural disease detection and recognition, complex factors such as the complexity of leaf disease causes, varying symptoms of the same disease at different stages, substantial differences within the same disease, minimal differences between different diseases, and environmental influences during the collection process make it difficult for CNN-based leaf disease detection methods to effectively focus on the diseased areas of the leaves. To address this, this paper proposes a grape leaf disease detection method combining CNN and Transformer structures. The research contributions are as follows:

1. To address the issue of insufficient samples for some categories of grape disease due to practical conditions, this paper uses the FastGAN adversarial generative network model to generate a large number of sample images, enriching the dataset of grape leaf diseases at different periods.

2. This paper proposes a lightweight neural network model, DLVTNet, which incorporates a novel Transformer structure, LVT Block, and a CNN structure, MARI Block, to process the global and local information of images respectively, and uses Dense connection methods to enhance the model's reuse of input information.

3. To achieve lightweight Transformer structures, this paper proposes a dimension-based lightweight self-attention mechanism, CLSHSA, which ensures recognition accuracy while achieving model lightweight. Additionally, this paper introduces a multi-scale attention mechanism, MELA, which combines positional and multi-scale information to obtain attention weights in images, aiding the model in effectively extracting diseased areas.

## 2. Materials and Methods

### 2.1. Image datasets and preprocessing

The grape leaf disease dataset used in this study comes from the publicly available plant disease classification dataset New Plant Diseases Dataset on Kaggle (https://www.kaggle.com/datasets/vipoooool/new-plant-diseases-dataset). Grape leaf images from this dataset were selected as the image dataset for this study. The dataset includes images of three types of grape leaf diseases as well as healthy leaves, totaling 7,222 images, divided into four categories. The images have been resized to 256×256 pixels and processed using oversampling. Sample images from the dataset are shown in Fig. 1.

(a) (b) (c) (d)

**Figure 1.** Grape leaf dataset sample examples. (a) Black Measles (b) Black rot (c) Leaf blight (d) Healthy

Deep learning neural network models typically require large datasets for training. This can provide diverse information for models, helping them to better learn the data distribution and reduce overfitting. In practical applications, models can perform more consistently when faced with various input information. However, traditional data augmentation methods such as rotation and mirroring can only change the spatial position of the original information and are difficult to effectively enrich the dataset information. In the detection of agricultural leaf diseases, the expression of disease characteristics can vary significantly due to different times of occurrence, making traditional data augmentation methods ineffective for enriching the information in leaf disease image datasets. Additionally, the significant differences in the number of images for different diseases in the dataset due to real-world environmental factors can lead to the model overlearning classes with more samples during training and misjudging classes with fewer samples. Therefore, this paper uses the FastGAN adversarial generative network model to enhance existing datasets, increasing the diversity of image dataset samples and balancing the dataset. The images obtained through different processing methods are shown in Fig. 2. Then, the augmented dataset is divided into training and testing sets in an 8:2 ratio, as shown in Table 1. Table 1 displays the number of samples in the grape leaf disease dataset used in this study, where “Raw dataset” represents the different categories of grape leaf images from the public New Plant Diseases Dataset, and oversampling was used with traditional data augmentation. “New dataset” represents the number of different samples generated using the FastGAN adversarial generative network. Finally, the number of image samples in the training and testing sets is included. In this study, to effectively observe the differences in the enhancement effects between the traditional data augmentation method and the use of the FastGAN adversarial generative network, the t-SNE algorithm was used for the visualization processing of the image dataset samples, with the results shown in Fig. 3. Fig. 3(a) and Fig. 3 (b) represent the visualization results of datasets obtained through two different image augmentation methods: traditional data augmentation and the use of the FastGAN adversarial generative network. In Fig. 3(a), the distribution of the four grape leaf categories is chaotic, with samples of different categories interwoven, indicating confusion in the dataset where image sample features are chaotic, and similar features are present in different leaf categories, but there is significant variation in features among similar samples. In Fig. 3(b), although there is also some distribution confusion, the four different leaf samples can gather around different areas, and similar feature expression patterns can be found among samples of the same category.

(a) (b) (c) (d)

**Figure 2.** Generation of different grape leaf categories using Fastgan. (a) Black Measles (b) Black rot (c) Leaf blight (d) Healthy

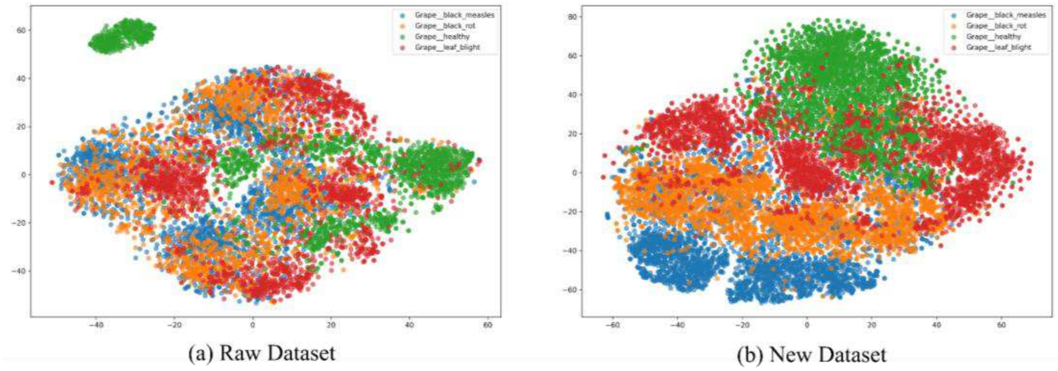



**Figure 3.** Visualization results of T-SNE on different data-augmented datasets. (a) stands for Raw dataset and (b) stands for New dataset

**Table 1.** Comparison of samples before and after data augmentation and the number of training and test sets

| | Black Measles | Black rot | Leaf blight | Healthy |
|---|---|---|---|---|
| Raw dataset | 1888 | 1920 | 1692 | 1722 |
| New dataset | 3000 | 3000 | 3000 | 3000 |
| train | 2400 | 2400 | 2400 | 2400 |
| test | 600 | 600 | 600 | 600 |

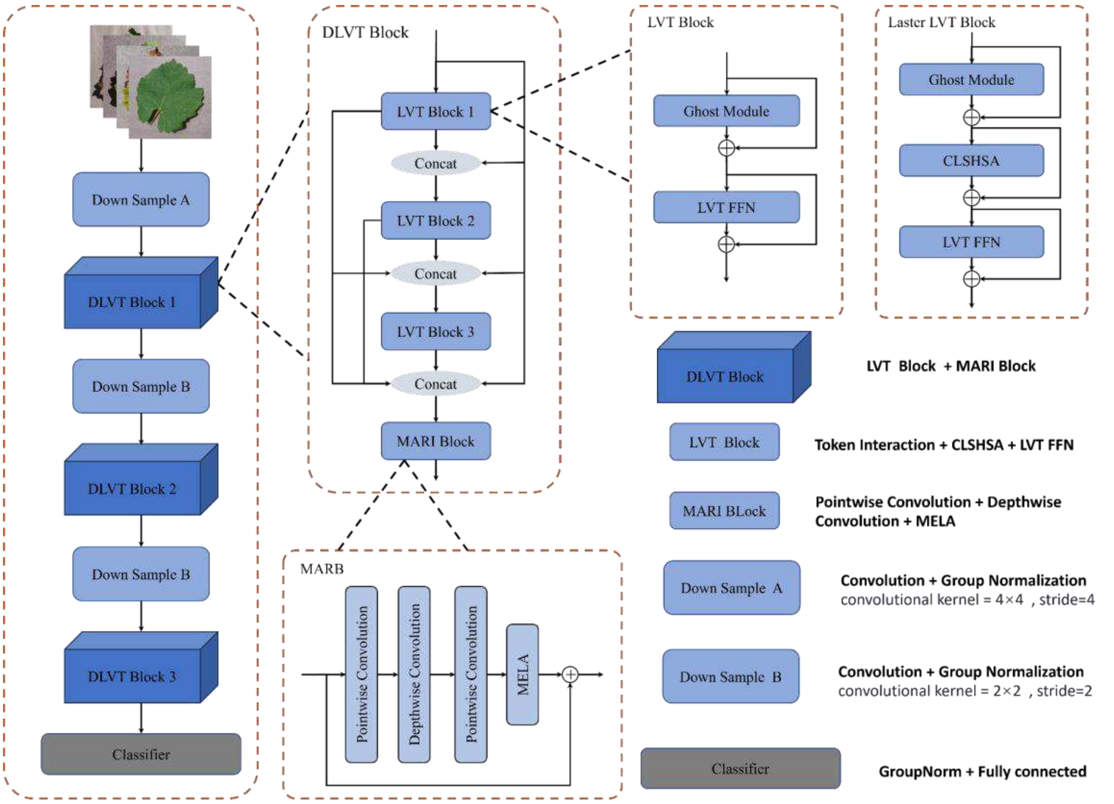



**Figure 4.** The main framework diagram of a lightweight neural network model that combines CNN and Transform

## 2.2. DLVTNet model body framework

The paper proposes a neural network model named DLVTNet, which combines CNNs and Transformers, primarily consisting of a DLVT Block (Dense-connected Light-weight Vision Transformer Block) and a Down Sample (downsampling module). This model is capable of recognizing and detecting diseases on grape leaves. As shown in Figure 4, the DLVT Block uses dense connections to connect the LVT Block (Lightweight Vision Transformer Block) and the MAIR Block (Multi-scale Attention Inverted Residual Block) to increase the reuse of input image information[32]. Additionally, within the LVT Block, the Ghost Module is used to extract multi-scale information from images, and the LVT FFN enhances the model's nonlinear capabilities. Furthermore, self-attention mechanisms are used only in the final LVT Block to reduce the computational resources required by the model. Within the DLVTNet model structure, the DLVT Block is primarily used to extract features from images, followed by downsampling operations performed by the DS Block structure. In this model, the first layer of the downsampling module uses a 4x4 convolutional layer to acquire spatial features from a larger receptive field. To avoid losing detailed features of the input image due to the use of large convolutional kernels, the second and third layers of the downsampling module use 2x2 convolutional layers for downsampling.

### 2.3. Dense-connected Lightweight Vision Transformer Block

In the detection of agricultural leaf diseases, the main difficulty lies in the complex and diverse representation of disease features, the issue of small and unevenly distributed disease features, which have posed challenges for the recognition of leaf diseases. In the detection of grape leaf diseases, the main challenges include the uneven spatial distribution of features, the unique expressions of leaf diseases at different stages, and the subtlety of early disease features. Therefore, this paper proposes a structure named DLVT Block that combines CNNs and Transformers. The DLVT Block primarily consists of the Transformer structure LVT Block and the CNN structure MAIR Block, and uses dense connections to concatenate each layer's output with the next layer's input along the channel dimension. As shown in Fig. 5, the DLVT Block structure includes three LVT Blocks and one MAIR Block. The first three LVT Blocks do not add self-attention mechanisms, and instead increase the expressiveness of images through the Ghost Module and LVT FFN. Then, through the LVT Block 3 with added self-attention mechanism, global information in the images is extracted. Finally, by using dense connection methods, the outputs of the first three LVT Blocks are concatenated with the DLVT Block's input along the channel dimension before the MAIR Block extracts local feature information from the input.

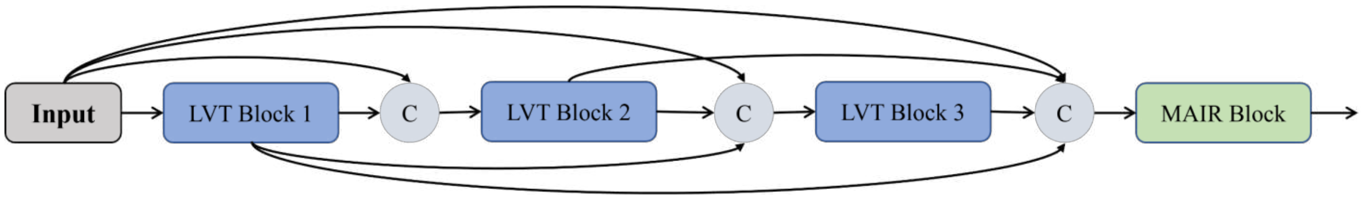



**Figure 5.** DLVT Block structure

#### 2.3.1. Lightweight Vision Transformer Blok

The Transformer architecture has achieved numerous advancements in the field of visual detection and has introduced many structures applicable to visual detection tasks. It mainly consists of self-attention mechanisms and Feed-Forward Network (FFN). The self-attention mechanism effectively extracts global information from the input, while the FFN enhances the expressive power of input image features.Among them, Swin Transformer uses the W-MSA module based on shifted window multi-head self-attention mechanism to replace the global attention mechanism in VIT, and then uses two layers of MLP with Gelu non-linear activation function in the middle to enrich the image features extracted by W-MSA, as shown in Fig. 6(a). To achieve a lightweight Transformer model, EfficientFormer proposes a new Attention variant called GroupAttention, which is combined with the deep learning-based Token Interaction structure and linear FFN (feedforward neural network) to form a new Transformer Block, as shown in Fig. 6(b).

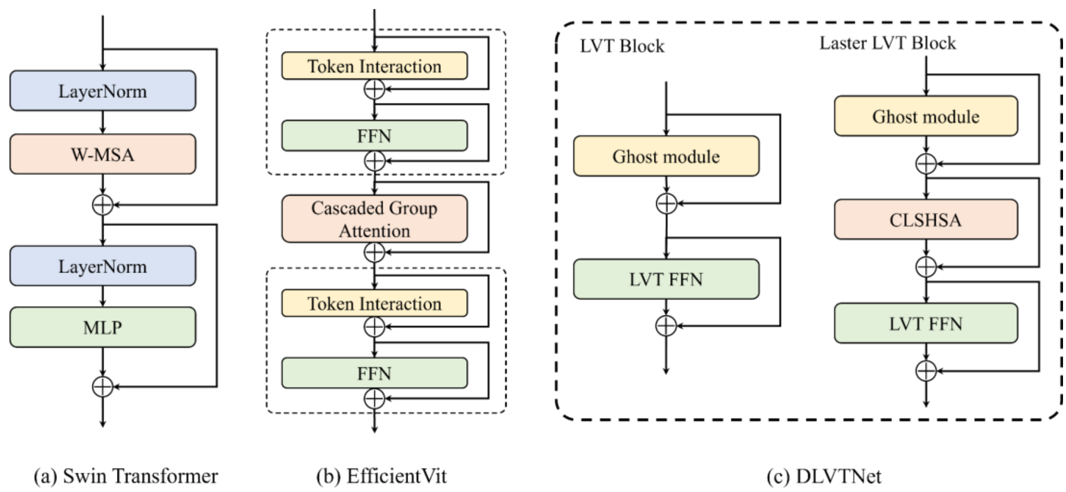



**Figure** 6. Different Vision Transformer architectures. (a) are Swin Transformer, (b) are EfficientVit, and (c) are DLVTNet

To further lightweight the Transformer model, this paper proposes a new Transformer Block structure, mainly consisting of Ghost model, CLSHSA, and LVT FFN, as shown in Fig. 6(c). Among them, Token Interaction is mainly responsible for local feature aggregation or conditional position embedding, CLSHSA is mainly responsible for modeling global context, and LVT FFN is mainly responsible for channel interaction. By combining Token Interaction and CLSHSA, local and global dependencies are captured in a lightweight manner. To reduce computational redundancy in the neural network model, CLSHSA is not used in the first and second LVT Blocks, but dense connections are used to input the first two layers of LVT Block into the last LVT Block to obtain global feature information from multi-level features of the input image, and then the obtained feature information is passed to the last layer of MARI Block. In the LVT block, there are two situations depending on the stage. In each DLVT Block, the first two LVT Blocks do not include CLSHSA, but instead use dense connections to uniformly input the output images obtained from the first two LVT Blocks into the last LVT Block with CLSHSA, which helps to better capture multi-level features of the input image.

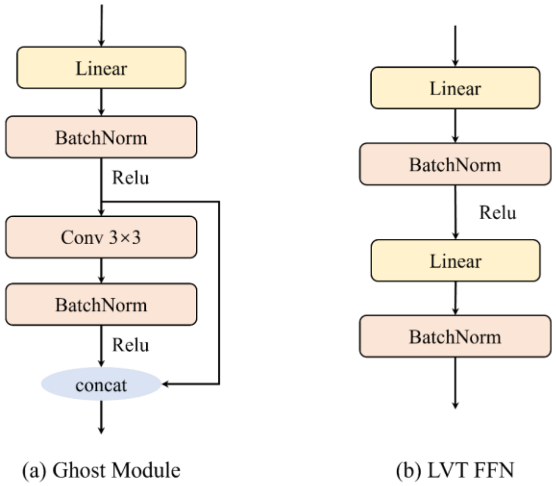



**Figure 7**. Structure diagram of Ghost module and LVT FFN. (a) is the Ghost Module and (b) is the LVT FFN

The Ghost module and LVT FFN structures are shown in Fig. 7. Fig. 7(a) shows the ghost module, which mainly processes input data at multiple scales to introduce more local inductive bias. In its structure, a 1×1 convolution is first used to compress the number of channels of the input image, then depth-separable convolution is used to generate more feature maps, and different feature maps are concatenated together to generate new outputs. This method can effectively reduce the number of parameters and computational cost required for calculation while generating rich feature maps.

LVT FFN mainly consists of two 1×1 convolutions and a Relu non-linear activation function, which can further non-linearly map and extract high-level features from the input feature maps, as shown in Fig. 7(b). First, when the feature map passes through the first layer of 1×1 convolution, it can expand the number of image channels. After passing through the non-linear activation function, it then passes through the second 1×1 convolution to integrate the image channels to retain important image feature information channels.

### 2.3.2. Channel Lightweight Single-Head Self-Attention

In the Transformer architecture, global information extraction from input data is primarily achieved through the self-attention mechanism. In self-attention, this is accomplished by calculating the relationships between each element in the input image and other elements to extract global feature information. Initially, the image features are processed through linear transformations to obtain three matrices: Q (Query), K (Key), and V (Value). The matrices K and Q are used to compute the attention weights, while V represents the values used for weighted summation to obtain the final output. However, due to the quadratic computational complexity of self-attention relative to the size of the image, neural network models using Transformer structures require substantial computational resources. As a result, many lightweight methods for the self-attention mechanism have been proposed.

Fig. 8 shows the design of various single-head attention mechanisms. Fig. 8(a) shows a single-head self-attention mechanism that retains all channels of the input image, then calculates the weights of the self-attention mechanism and weights the values in the image. However, this method considers all image channels when calculating K and Q, resulting in some computational redundancy. Fig. 8(b) shows a lightweight single-head self-attention mechanism proposed in the SHViT model. In this attention mechanism, a pre-convolution method is used to split the input image channels into two parts. One part of the channel images uses a self-attention mechanism to obtain global information. The other part of the channel images is not processed. Finally, the feature maps obtained from both parts are concatenated.

In this paper, to achieve a lightweight Transformer structure, a single-head self-attention mechanism combined with CNN channel attention, called CLSHSA, is proposed. The inspiration primarily comes from the fact that the matrices Q and K in self-attention mechanism calculations generate significant redundancy, and the channel dependency of matrices Q and K is lower than that of matrix V when processing images[33]. The CLSHSA structure is shown in Fig. 9(a). Before calculating the matrices K and Q, the input image has its channel dimensions reduced using a CL Block (Channel Lightweight Block). Self-Attention is then used to weight the input matrix V, thereby extracting global information and achieving a lightweight processing of the self-attention mechanism[34]. The computation of CLSHSA is illustrated in Eq. 1, Eq. 2, and Eq. 3.

In the CL Block, the main components are a fully connected layer, pooling layer, Sigmoid activation function, and ReLU activation function, as shown in Fig. 9(b). The data input to the CL Block undergoes global pooling first, followed by two fully connected layers and the ReLU activation function to obtain the channel weight information of the image. The Sigmoid function is then used to generate a weight vector to apply weighting to the image channels. Finally, another fully connected layer integrates the weighted channel information and removes redundant channel information, achieving lightweight processing of the self-attention mechanism.

$$X_r = \mathrm{CL}(X) \tag{1}$$

$$\mathrm{Attention}(Q_r, K_r, V) = \mathrm{Softmax}\left(\frac{Q_r K_r^T}{\sqrt{d_r}}\right)V \tag{2}$$

$$\mathrm{CLSHSA}(X) = \mathrm{Attention}(XW^Q, XW^K, XW^V) \tag{3}$$

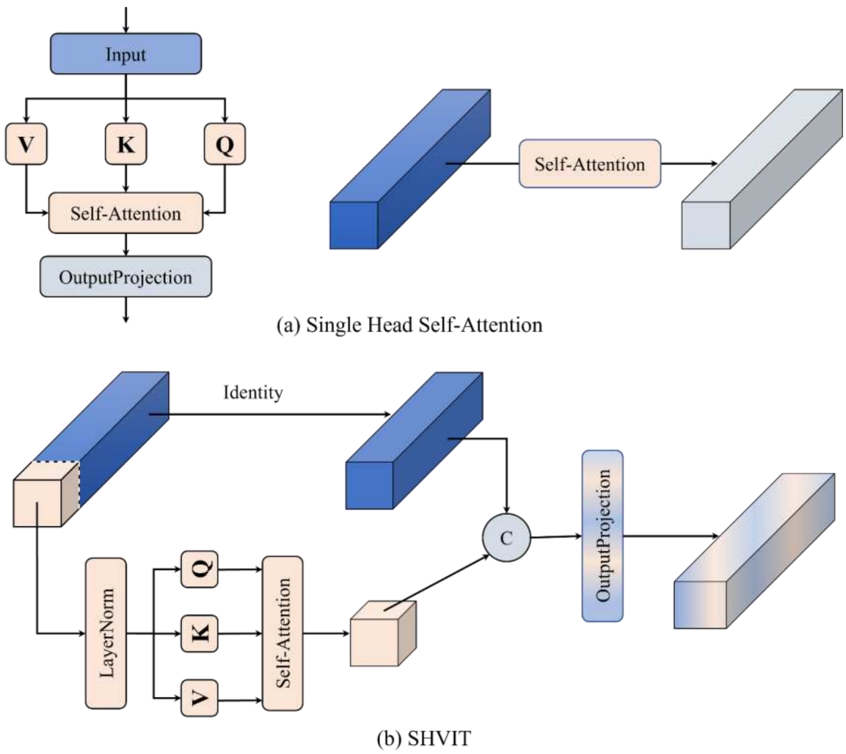



**Figure 8**. Different single-head attention mechanisms and lightweight methods (a) are traditional single-head self-attention mechanisms and (b) are SHVIT's improved single-head self-attention mechanisms

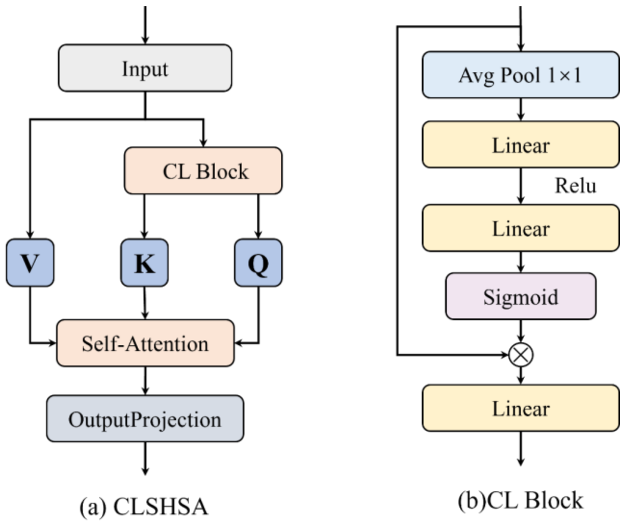



**Figure 9**. Structure of CLSHSA attention mechanism. (a) is the CLSHSA and (b) is the CL Block

### 2.4. Multi-scale Attention Inverted Residual Block

In this paper, to effectively extract features of grape leaf diseases, a DLVT Block combining CNN and Transformer structures is proposed. To achieve effective extraction of local feature information, a Multi-scale Attention Inverted Residual Block (MAIR

Block) with a multi-scale attention mechanism is introduced, as shown in Fig. 10. This structure uses an inverted residual block design, which allows for feature extraction from images without significantly increasing the model's computational resource requirements. The structure is illustrated in Fig. 10(a) and mainly consists of two 1×1 pointwise convolutions, one 3×3 depthwise convolution, Batch Normalization, and the Multi-scale Efficient Local Attention (MELA) mechanism. The input image first undergoes a 1×1 pointwise convolution to expand the image's expressiveness, then the expanded channels are processed with depthwise convolution, and ReLU activation function is applied to enhance the non-linear expression of image features. Finally, a second pointwise convolution integrates the number of channels, and the integrated image is fed into the MELA attention mechanism.

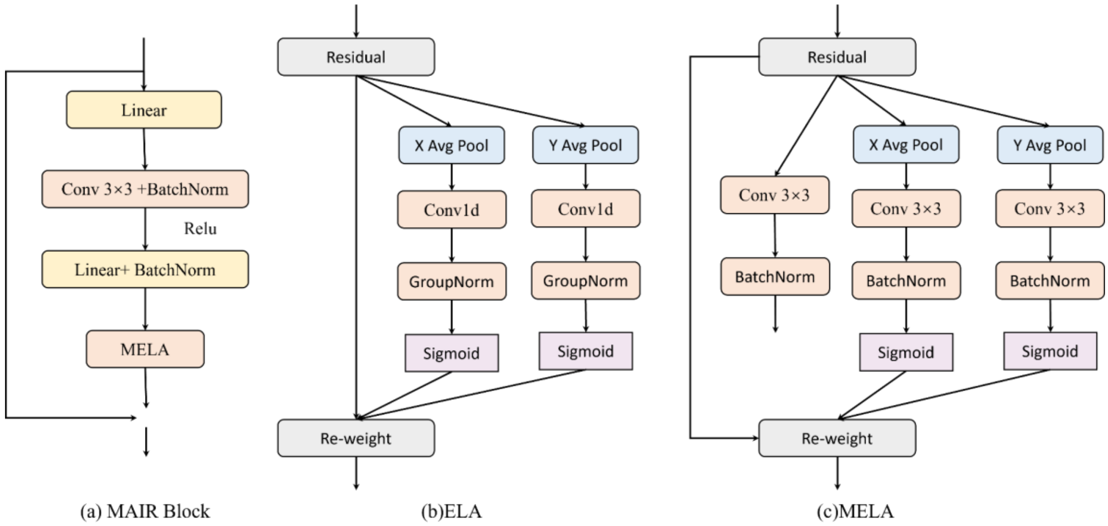



**Figure 10.** Schematic diagram of the structure of the MAIR Block. (a) is the MAIR Block, (b) is the ELA attention mechanism, and (c) is the MELA attention mechanism

In this paper, to effectively extract local feature information from images after concatenation, a Multi-scale Efficient Local Attention (MELA) mechanism is proposed. MELA is inspired by the CA (Coordinate Attention) and ELA (Efficient Local Attention) mechanisms. It obtains spatial weights for different regions of the image based on horizontal and spatial positional information of the input image[35]. ELA, utilizing the coordinate positional information approach from CA, redesigns the attention mechanism structure to achieve a lightweight implementation without dimensionality reduction. The structure is shown in Fig. 10(b). The goal of this paper is to achieve effective extraction of features with uneven spatial distribution. To this end, a multi-scale attention mechanism, MELA, is constructed, with its structure illustrated in Fig. 10(c).

The image input to the MELA attention mechanism is processed in two parts. One part uses average pooling with kernels of (H, 1) and (1, W) to extract horizontal and vertical feature vectors, respectively. These extracted feature vectors are then processed using a 3×3 convolutional layer and Batch Normalization, with Sigmoid calculating the spatial weights of the input image in the horizontal and vertical directions. In the other part, the input image is directly processed using a 3×3 convolutional layer and Batch Normalization, followed by Sigmoid to obtain the spatial weights of the image. These weights are combined with the horizontal and vertical weights of the image to perform weighted processing. This method effectively focuses on disease areas in the image by combining the weights at different locations and the multi-scale weights of the entire image.

In the MELA attention mechanism, to obtain attention weights at different scales, the input representations of height h and width w in channel c are first obtained using pooling kernels of (H, 1) and (1, W) along the vertical and horizontal directions, respectively. The calculation process is shown in Eq. 4 and Eq. 5.

$$z_c^h(h)=\frac{1}{W}\sum_{0\le i<h} x_c(h,i) \tag{4}$$

$$z_c^w(w)=\frac{1}{H}\sum_{0\le i<w} x_c(h,w) \tag{5}$$

Where $z_c$ represents the encoding result of average pooling in the c-th channel in the horizontal direction w and vertical direction h; $x_c$ represents the feature value of the c-th channel at height h and width w in the feature map. To effectively utilize the obtained feature vectors, we apply one-dimensional convolution (conv1d) to process information in the horizontal and vertical directions. The processed information is then enhanced using Batch Normalization (BN), and finally, the attention weights are computed using the Sigmoid function. The calculation process is shown in Eq. 6 and Eq. 7. In these equations, $F_w$ and $F_h$ represent the convolution operations with 3×3 kernels in the horizontal and vertical directions, respectively.

$$g^h=\mathrm{sigmoid}\left(\mathrm{B_h}\left(\mathrm{F_h}\left(z^h\right)\right)\right) \tag{6}$$

$$g^w=\mathrm{sigmoid}\left(\mathrm{B_w}\left(\mathrm{F_w}\left(z^w\right)\right)\right) \tag{7}$$

Then, to obtain multi-scale global weights, the input image feature information is processed using a convolution operation $F_2$ with a 3×3 kernel and Batch Normalization $B_2$, followed by the calculation of weights using the Sigmoid function. The calculation method is shown in Eq. 8.

$$g^s=\mathrm{sigmoid}\left(\mathrm{B_2}\left(\mathrm{F_2}\left(x\right)\right)\right) \tag{8}$$

Finally, the obtained weights $g^h$ and $g^w$ and multi-scale weights $g^s$ are used to weight the input image, resulting shown in Eq. 9.

$$y(i,j)=x_c(i,j)\times g_c^h(i,j)\times g_c^w(i,j)\times g^s(i,j) \tag{9}$$

Where $y$ is the final output of MELA. Inspired by CA attention and ELA attention, MELA can effectively combine spatial position information and multi-scale information to find regions of interest in the image, improving accuracy.

## 3. Experimental results

### 3.1. Experimental design

The computer used in this paper runs on Windows 11 operating system, using a 12th Gen Intel(R) Core(TM) i7-12700 (2.10 GHz) processor. GPU acceleration is used for model training and testing, with an NVIDIA GeForce RTX 3060 (12G) GPU. The software environment uses Python 3.9.13, PyTorch 1.13.1, and CUDA 11.6 framework. Additionally, in training the neural network model, Adam is set as the optimizer, the learning rate is set to 0.000001, the number of iterations is set to epochs = 200, and the batch size is set to 16.

### 3.2. Evaluation indicators

In this study, precision, recall, accuracy, and F1 score are used to evaluate the recognition effect of different models on the grape leaf dataset. The calculation methods are shown in Eq. 10, Eq. 11, Eq. 12, and Eq. 13.

$$Precision=\frac{TP}{TP+FP} \tag{10}$$

$$Recall=\frac{TP}{TP+FN} \tag{11}$$

$$Accuray=\frac{TP+TN}{TP+FN+FP+TN} \tag{12}$$

$$F1=\frac{2TP}{2TP+FP+FN} \tag{13}$$

Where TP (True Positive) represents the number of positive samples correctly predicted as positive by the model; TN (True Negative) represents the number of negative samples correctly predicted as negative by the model; FP (False Positive) represents the number of negative samples incorrectly predicted as positive by the model; FN (False Negative) represents the number of positive samples incorrectly predicted as negative by the model.

In this study, precision represents the proportion of samples correctly judged as positive by the network model. Recall measures the proportion of positive samples correctly identified by the network model among the actual positive samples. Accuracy represents the proportion of total samples correctly classified by the network model. The F1 score comprehensively considers precision and recall, and is the harmonic mean of precision and recall. In addition, two parameters, Flops (floating-point operations) and Params (number of parameters), are introduced to evaluate the size of the model. The larger these two parameters, the more computational resources the model requires.

### 3.3. Comparative experiments of different network models

In this study, to test the recognition performance of the DLVT model, comparison experiments were conducted using existing mainstream image classification models. The comparative models discussed in this paper mainly include CNN models such as ConvNext[36], EfficientNet[37], MobileNet[38], ResNet[39], DenseNet[40], and InceptionNext[41], as well as Transformer models such as Deit[42], EfficientFormer[43], MobileVit[44], Swin Transformer[45], and TinyVit[46]. The evaluation results of the comparative experiments are shown in Table 2. Among the 12 network models participating in the experiment, six models achieved an average recognition accuracy of over 96%: EfficientFormer, Swin Transformer, ResNet, DenseNet, TinyVit, and DLVTNet. Among these, DenseNet and DLVTNet achieved average recognition accuracies of over 98%.

**Table 2.** Comparison of average Accuracy, Pression, Recall, F1 score, Flop and Params of different network models

| Methods | Accuracy | Precision | Recall | F1 score | Flop(G) | Params(M) |
|---|---|---|---|---|---|---|
| ConvNext V2 | 95.89 | 96.02 | 95.98 | 95.97 | 4.45 | 27.79` |
| Deit3 | 95.95 | 95.92 | 95.89 | 95.82 | 4.24 | 21.97 |
| EfficientNet V2 | 95.73 | 95.59 | 95.52 | 95.39 | 2.85 | 21.305 |
| EfficientFormer V2 | 96.82 | 96.11 | 96.09 | 96.06 | 1.23 | 12.637 |
| Mobilenet V3 | 88.16 | 88.14 | 88.05 | 87.50 | 0.2 | 4.189 |
| MobileVit V2 | 93.45 | 92.64 | 92.58 | 92.55 | 1.41 | 4.87 |
| SwinTransformer V2 | 96.29 | 96.29 | 96.23 | 96.18 | 4.51 | 28.33 |
| ResNet18 | 97.71 | 97.70 | 97.69 | 97.65 | 1.82 | 11.69 |
| DenseNet | 98.30 | 98.06 | 98.06 | 98.04 | 2.83 | 7.89 |
| InceptionNext | 93.81 | 92.75 | 92.73 | 92.72 | 4.20 | 28.04 |
| TinyVit | 96.20 | 95.54 | 95.43 | 95.46 | 1.19 | 12.07 |
| DLVTNet | **98.48** | **98.48** | **98.47** | **98.46** | 0.49 | 1.05 |

Table 2 also reflects the Flops and Params required for the different network models in the comparison experiment, which can be used to evaluate the computational resources required by each model. Of the 12 neural network models participating in the comparison experiment, only MobileNet V3 has a smaller Flops requirement than DLVTNet. However, MobileNet V3's Params requirement is four times that of DLVTNet, and its average recognition accuracy is only 88.16%. DLVTNet requires only 0.49G of Flops, making it the second smallest after MobileNet, and it has the smallest Params requirement among the models tested, with a maximum difference of up to 28 times compared to others. Despite this, DLVTNet has the highest average recognition accuracy and Recall, with parameters of 98.48, 98.48, 98.47, and 98.46, respectively, as highlighted in black and bold in Table 2. Fig. 11 shows the accuracy curve comparison of the network models in the parameter comparison experiment. Fig. 11(a) compares the accuracy curves of DLVTNet and CNN models, while Fig. 11(b) compares those of LVTNet and Transformer models, marking the peak recognition accuracy curve. As shown in Fig. 11(a), the recognition accuracy of the DLVTNet model can reach up to 99.79%, and compared to other neural network models, the DLVTNet model's accuracy curve is stable and converges to a higher accuracy.

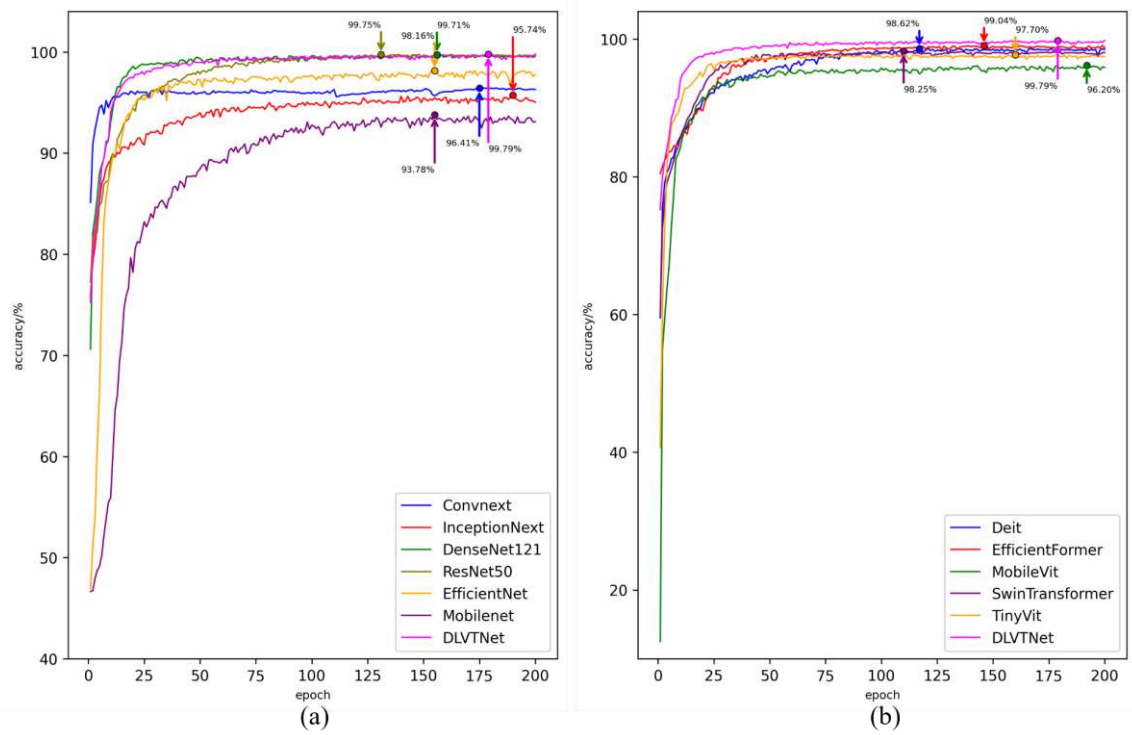



**Figure 11**. Training accuracy curves of different models (a) is a comparison of CNN models and (b) is a comparison of Transformer models

In addition, to effectively express the recognition accuracy of the LDVTNet model for grape diseases, this paper also conducted confusion matrix tests on the models appearing in Table 2, and the test results are shown in Fig. 12. In the confusion matrix, B M, B R, H, and L B correspond to black_measles, black_rot, healthy, and leaf_blight, respectively, which are the four categories of grape leaves. These are abbreviations of the first letters of the grape diseases shown in Fig. 1. In the detection results shown in Fig. 11, the LDVTNet model achieved correct recognition of all sample images, while DenseNet, ResNet, Swin Transformer, Tiny Vit, and EfficientFormer performed slightly worse, with recognition errors of less than 5 samples. The remaining models have larger errors, with Deit3, MobileNet, and InceptionNext having more than 10 misrecognized samples, mainly distributed in the Black Measles and Black Rot disease categories. This problem is mainly due to the similarity in spatial distribution and color features of these two disease characteristics in the early stages of the disease, making it difficult for network models to effectively distinguish between them.

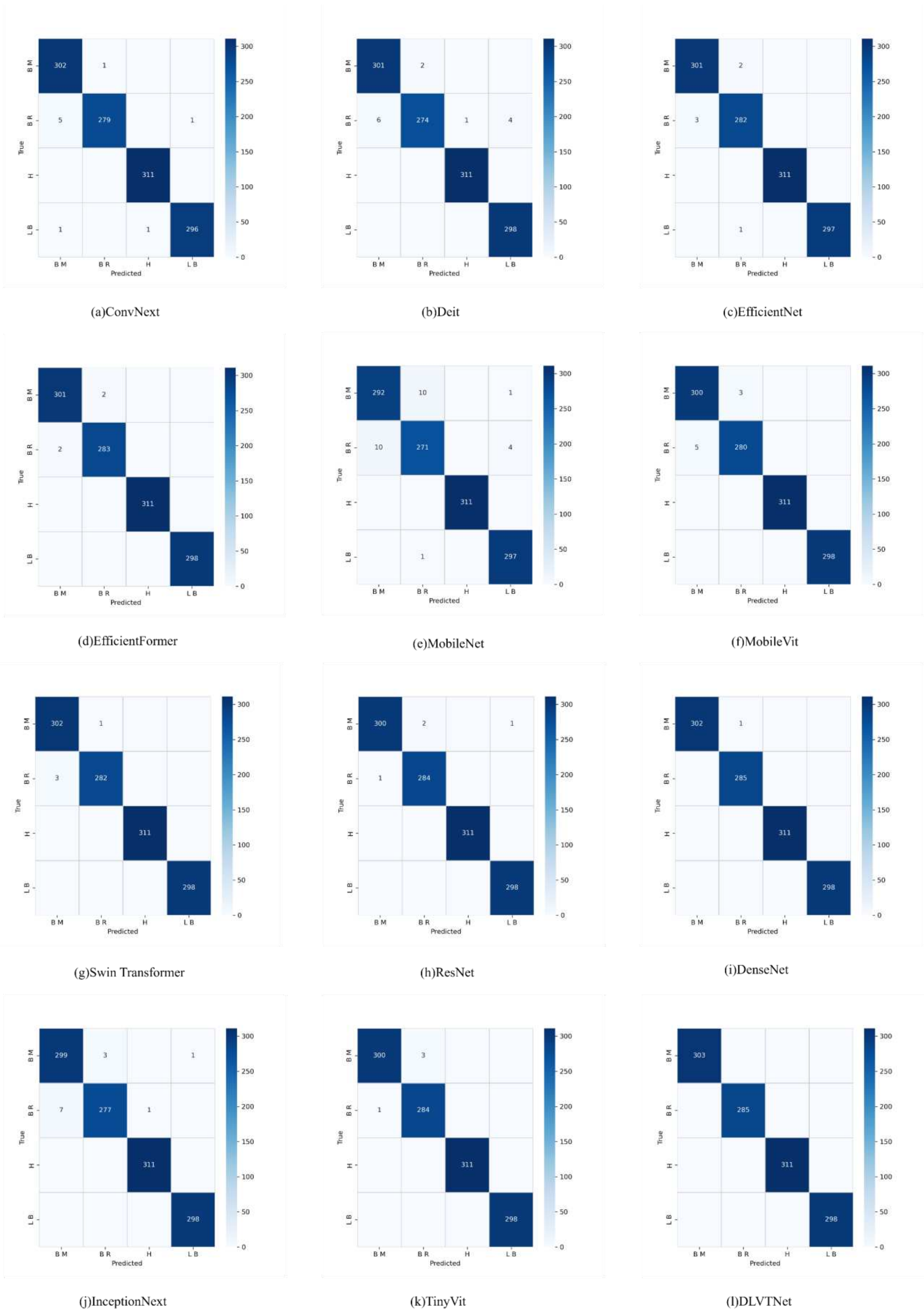



**Figure 12.** Confusion matrix test renderings for different network models, (a) to (l) are the network models shown in Table 2, respectively

### 3.4. Ablation experiments

In this paper, ablation experiments were also conducted to test the impact of different methods on the recognition performance of the DLVTNet model. During these ablation experiments, six distinct approaches were examined: using the original grape leaf dataset, using the dataset enhanced by FastGAN, adding the Ghost Module, adding CLSHSA, adding the MARI Block, and adding Dense connections. Table 3 reflects parameters such as average recognition accuracy, Precision, Recall, and F1 score obtained through various training methods. Fig. 13 illustrates the clustering effects of different models on grape leaf disease identification using the t-SNE algorithm for visualization. Among the six methods tested in the ablation experiments, the use of the FastGAN data-enhanced image dataset yielded the most significant improvement in model recognition ability, elevating the average recognition accuracy from 86.76% to 94.30%. Furthermore, the clustering effects of these two methods on grape leaf disease samples are depicted in Fig. 13(a) and Fig. 13(b). Fig. 13(a) shows the results from training with the original dataset, where samples from different categories are mixed, making it challenging to achieve effective clustering of different disease types.

**Table 3.** Data diagram of ablation experiment shows the effect of different functions on the recognition ability of DLVTNet

| Method | Accuracy | Precision | Recall | F1 score | Flop(G) | Params(M) |
|---|---|---|---|---|---|---|
| Base | 86.76 | 86.23 | 86.45 | 86.23 | 0.329 | 0.719 |
| +GAN | 94.30 | 93.68 | 93.70 | 93.61 | 0.329 | 0.719 |
| +Ghost | 96.63 | 96.15 | 96.15 | 96.12 | 0.317 | 0.690 |
| +CLSHSA | 96.31 | 96.02 | 95.97 | 95.93 | 0.298 | 0.645 |
| +MARI | 97.92 | 97.53 | 97.52 | 97.51 | 0.406 | 0.860 |
| +Dense | **98.48** | **98.48** | **98.47** | **98.46** | 0.493 | 1.054 |

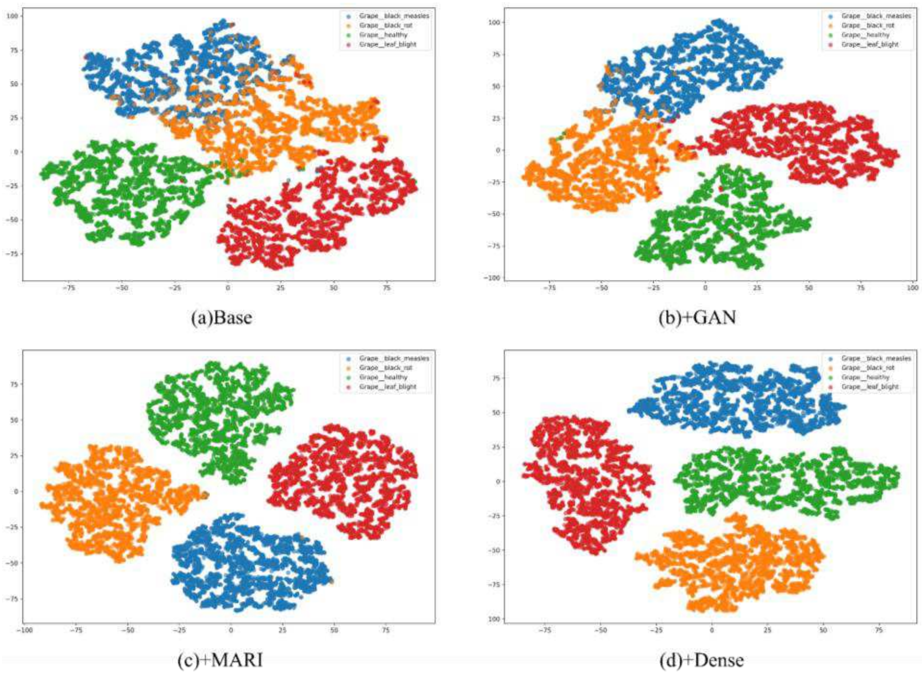


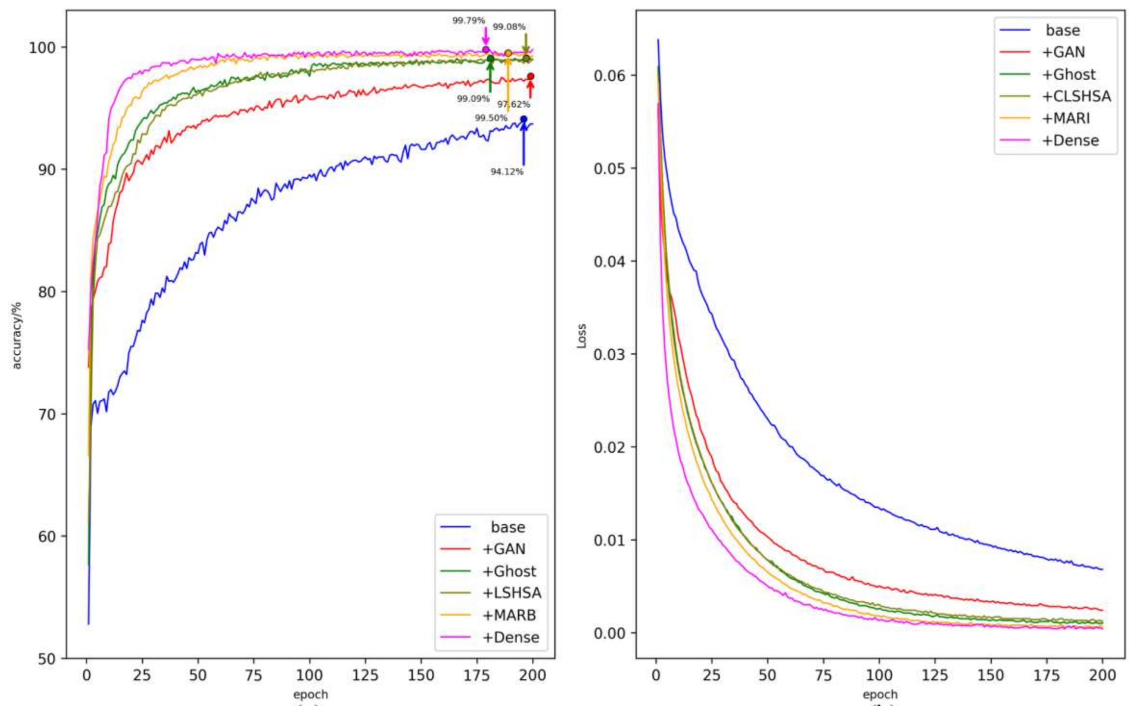

**Figure 13.** Ablation experiments used different methods of DLVTNet model to visualize the clustering effect of grape leaf categories

On the other hand, Fig. 13(b) demonstrates the visualization outcome after employing FastGAN data augmentation followed by training with the same framework structure. Although there are still a few misclassified samples among the four distinct categories, a certain degree of clustering effect is achieved. Additionally, this paper proposes the implementation of lightweight processing for the Transformer structure by adding the Ghost Module and CLSHSA. As evident from Table 3, the model's demand for computational resources is notably reduced, while its recognition capability is effectively enhanced, proving the efficacy of the lightweight approach presented in this paper. Moreover, the introduction of the MARI Block and Dense connections to the DLVTNet has increased the average recognition accuracy by 1.61% and 0.56%, respectively. As shown in Fig. 13(c), the incorporation of the

MARI Block into the DLVTNet network model enables effective clustering of four different grape leaf diseases, grouping samples from different categories into distinct regions, albeit with a few samples not clustered into the correct categories. By adding Dense connections, DLVTNet facilitates the reuse of input image information, efficiently extracting image feature information. The clustering effect for different disease samples is exhibited in Fig. 13(d), where virtually no samples are misclassified, and all tested images are categorized into their respective regions.

Fig. 14 compares the accuracy and loss value curves of the DLVTNet model using different methods in the ablation experiments. As shown in Fig. 14(a), the accuracy curve of the DLVTNet model incorporating the innovative methods presented in this paper demonstrates a significant advantage over other incomplete approaches. The accuracy of this model increases steadily, exhibits its minimal fluctuations, and achieves the highest recognition accuracy. Additionally, Fig. 14(b) shows the loss value change curves for different training methods. The loss value obtained from training with the original dataset decreases the slowest and is the highest among the six methods. With the addition of various improvement methods, the rate of decline in loss value continuously improves, with the loss value of the DLVTNet model incorporating all methods being optimal.

Furthermore, to visually observe the changes in input images using different DLVTNet models, feature maps from different DLVT Blocks within the model are visualized, as shown in Fig. 15. The figure primarily demonstrates the feature extraction effects of different network model frameworks on input images after FastGAN data augmentation. Among the five methods displayed, +GAN, +Ghost, +CLSHSA, and +MARI methods can abstractly preserve the leaf area in the input images, but they struggle to localize the disease within the leaf area, failing to retain the diseased regions in the feature extraction process. In contrast, the DLVTNet model with Dense connections effectively extracts and retains features of the diseased regions in the feature map output at the final stage.




(a) (b)

**Figure 14.** The accuracy of the ablation experiment is graphed, and the highest precision value of the curve is marked by the arrows. (a) is a comparison chart of the accuracy curve, and (b) is a comparison chart of the Loss value curve

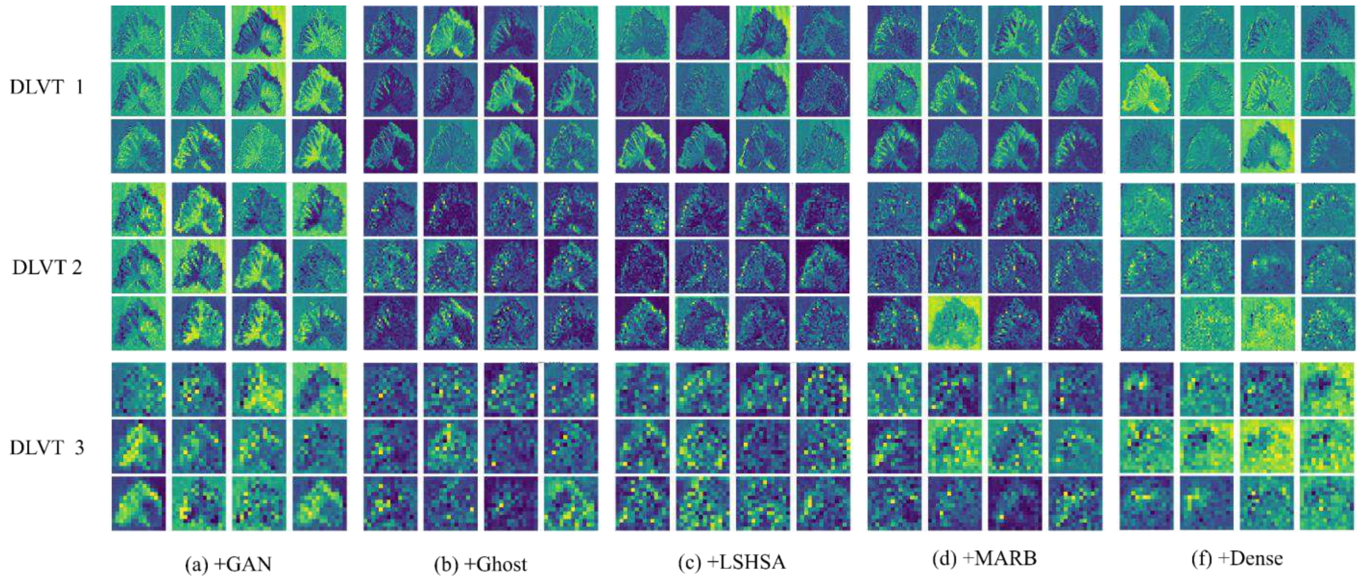



**Figure 15.** Characteristic plots of different DLVT block outputs of DLVTNet models using different methods

### 3.5. Comparative experiments with different attention mechanisms

This paper proposes a deep learning network model DLVTNet that combines CNN and Transformer. In this model, MELA is used as the last layer of the dense connection block to extract regions of interest from the input image. To verify the performance of the MELA attention mechanism, this section compares it with five mainstream and new attention mechanisms: ECA[47], CBAM[48], GAM[49], NAM[50], and ELA[51]. The experimental data obtained are shown in Table 4. Among them, ours has the highest average recognition accuracy, Precision, Recall, and F1 score, which are 98.48%, 98.48 %, 98.47%, and 98.46, respectively, about 0.63%, 0.33%, 1.98%, 0.31%, and 0.47% higher than the average accuracy of the other five attention mechanisms. The accuracy change curves of the six attention mechanisms are shown in Fig. 16, where the MELA attention mechanism can effectively improve the recognition accuracy of the model, with its highest accuracy reaching 99.79%, surpassing other mainstream attention mechanisms.

**Table 4.** Evaluation and comparison of training data of different attention mechanisms with ECA, CBAM, GAM, NAM, and ELA

| Methods | Accuracy | Precision | Recall | F1 score | Flop(G) | Params(M) |
|---|---|---|---|---|---|---|
| ECA | 97.85 | 97.86 | 97.90 | 97.86 | 0.491 | 1.051 |
| CBAM | 98.15 | 98.25 | 98.18 | 98.15 | 0.491 | 1.054 |
| GAM | 96.50 | 96.65 | 96.56 | 96.40 | 0.732 | 1.590 |
| NAM | 98.17 | 98.23 | 98.21 | 98.18 | 0.491 | 1.051 |
| ELA | 98.01 | 98.11 | 98.04 | 97.98 | 0.493 | 1.054 |
| MELA | **98.48** | **98.48** | **98.47** | **98.46** | 0.493 | 1.054 |

In addition, this study also visualizes the regions of interest in disease images for different attention mechanisms, and the results are shown in Fig. 17. It can be observed that using the MELA attention mechanism can effectively capture the leaf disease areas, with the best focus on disease areas among the six attention mechanisms. Furthermore, ECA, CBAM, and NAM attention mechanisms can also capture regions of interest, but their focus areas are relatively scattered and fail to accurately focus on the disease areas in the leaves, with more useless areas. Among the six attention mechanisms, GAM has the lowest accuracy, and its class activation map also has the worst effect on focusing on disease areas, failing to effectively localize diseases.

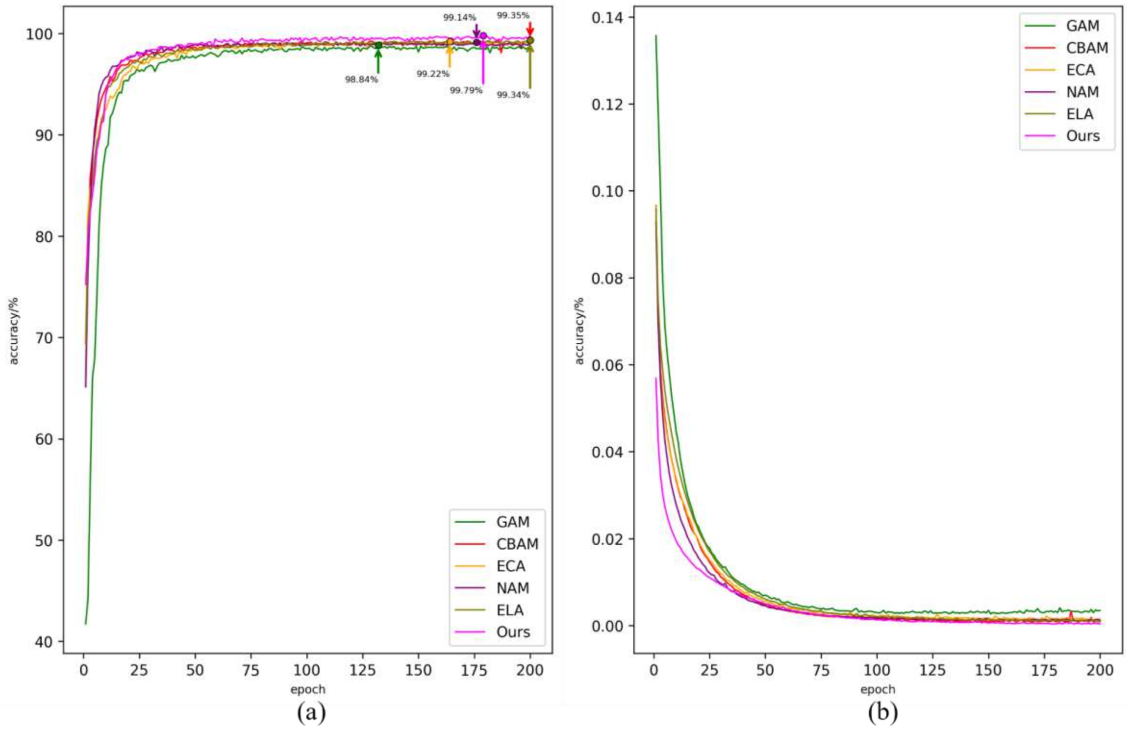



**Figure 16.** Comparison of attention mechanism model curves. (a) is a comparison chart of the accuracy curve, and (b) is a comparison chart of the Loss value curve

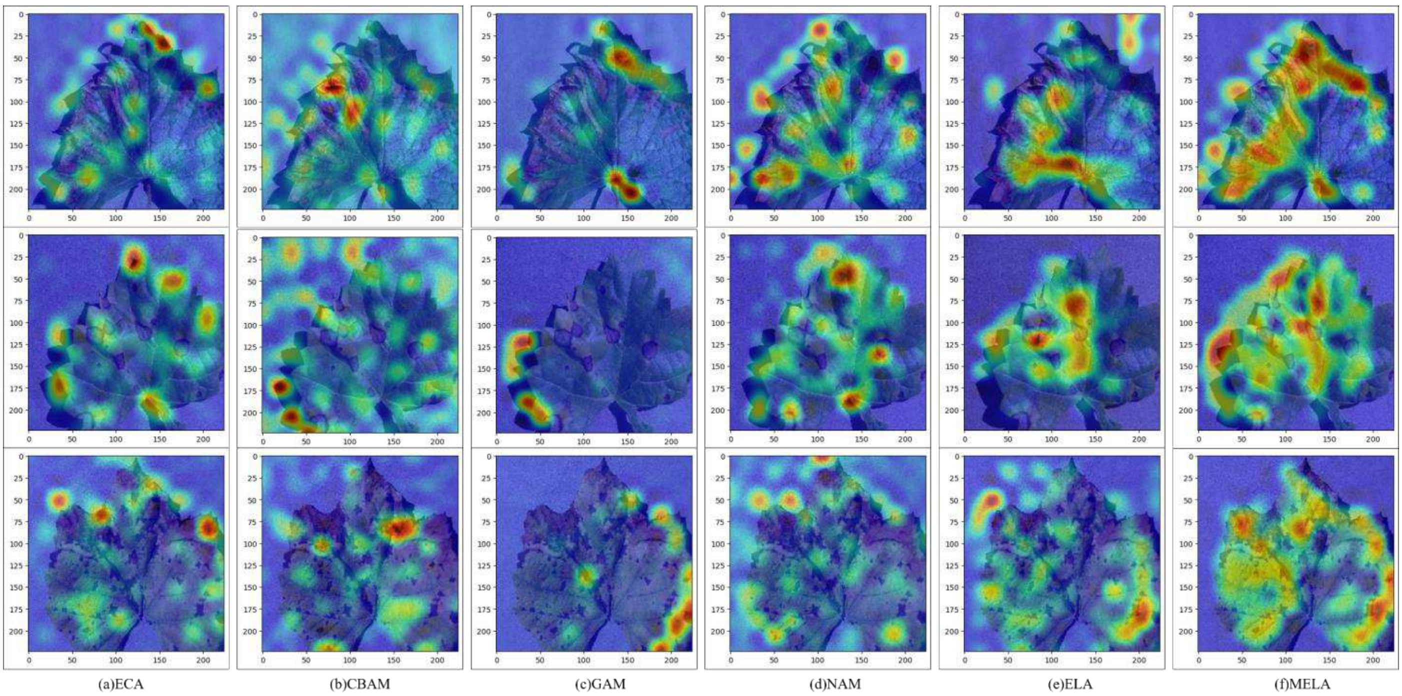



**Figure 17.** Category activation diagram of DLVTNet using different attention mechanisms. Mechanisms of attention in Table 5 are from (a) to (e), respectively. Each of the three rows represents a different disease category

### 3.6. Comparison experiments with different datasets

The method proposed in this paper demonstrates effective detection performance for grape leaf disease recognition and detection. However, using only a single type of leaf disease makes it difficult to prove the generalization capability of the DLVTNet model. To address this, this section tests the DLVTNet model across datasets using tomato leaf disease images from the publicly available New Plant Diseases Dataset (https://www.kaggle.com/datasets/vipoooool/new-plant-diseases-dataset). Different categories of tomato image samples are shown in Fig. 18. This dataset includes 10 types of leaf categories, with a total of 18,345 images. The training and test set splits are shown in Table 5.

**Table 5.** Classification of the training test set of the tomato disease dataset

| Categories | Dataset | Train | test |
|---|---|---|---|
| Bacterial spot | 1702 | 1362 | 340 |
| Early blight | 1920 | 1536 | 384 |
| Healthy | 1926 | 1541 | 385 |
| Late blight | 1851 | 1481 | 370 |
| Leaf mold | 1882 | 1506 | 376 |
| Septoria leaf spot | 1745 | 1396 | 349 |
| Spider mites | 1741 | 1393 | 348 |
| Target spot | 1827 | 1462 | 365 |
| Mosaic | 1790 | 1432 | 358 |
| Yellow leaf curl | 1961 | 1569 | 392 |

**Table 6.** Comparison of training parameters of different models in tomato disease datasets

| Methods | Accuracy | Precision | Recall | F1 score | Flop(G) | Params(M) |
|---|---|---|---|---|---|---|
| ConvNext V2 | 90.89 | 91.07 | 90.96 | 90.95 | 4.45 | 27.79` |
| Deit3 | 97.79 | 97.83 | 97.78 | 97.78 | 4.24 | 21.97 |
| EfficientNet V2 | 90.50 | 90.48 | 90.46 | 90.37 | 2.85 | 21.30 |
| EfficientFormer V2 | 97.98 | 98.02 | 97.97 | 97.94 | 1.23 | 12.63 |
| Mobilenet V3 | 89.12 | 89.11 | 89.08 | 88.97 | 0.2 | 4.18 |
| MobileVit V2 | 88.18 | 88.18 | 88.02 | 87.98 | 1.41 | 4.87 |
| SwinTransformer V2 | 96.78 | 96.89 | 96.80 | 96.79 | 4.51 | 28.33 |
| ResNet18 | 97.85 | 97.96 | 97.84 | 97.81 | 1.82 | 11.69 |
| DenseNet | 98.50 | 98.52 | 98.49 | 98.49 | 2.83 | 7.89 |
| InceptionNext | 94.03 | 94.18 | 94.07 | 94.04 | 4.20 | 28.04 |
| TinyVit | 95.04 | 95.09 | 95.00 | 94.96 | 1.19 | 12.07 |
| DLVTNet | **98.57** | **98.74** | **98.57** | **98.55** | 0.49 | 1.05 |

To compare with existing mainstream models, this section's experiments used models consistent with those in the comparative experiments of Part 3.3. These models include CNN models such as ConvNext, EfficientNet, MobileNet, ResNet, DenseNet, and InceptionNext, as well as Transformer models including Deit, EfficientFormer, MobileVit, Swin Transformer, TinyVit, and the DLVTNet proposed in this paper, totaling 12 network models. Table 6 presents the evaluation parameters for the 12 models trained on the tomato leaf dataset. Models with an average recognition accuracy exceeding 97% include Deit, EfficientFormer, ResNet, DenseNet, and DLVTNet. Among these, DenseNet and DLVTNet models have an average recognition accuracy above 98%, with DLVTNet achieving the highest average evaluation parameters among the 12 models, specifically 98.57%, 98.74%, 98.57%, and 98.55%.

Fig. 19 shows the accuracy curves for the 12 network models tested. Fig. 19(a) compares the accuracy curves of DLVTNet and CNN models. It shows that the accuracy curves of DLVTNet and ConvNext models experienced significant fluctuations in the early stages of training, but DLVTNet was able to converge to a higher level of accuracy. Fig. 19(b) compares the accuracy curves of the DLVTNet model with Transformer models. Most models, except for TinyVit and MobileVit, experienced significant fluctuations, but TinyVit and MobileVit had recognition accuracies much lower than the other models. In the accuracy curve comparison of different network models in Fig. 19, DLVTNet experienced significant fluctuations in the early training stages but exhibited a higher accuracy improvement trend compared to other models, with a peak recognition accuracy of 99.92%, the highest among the

12 neural network models. This indicates that while the DLVTNet model proposed in this paper has some limitations, it still achieves effective recognition of tomato leaf diseases.

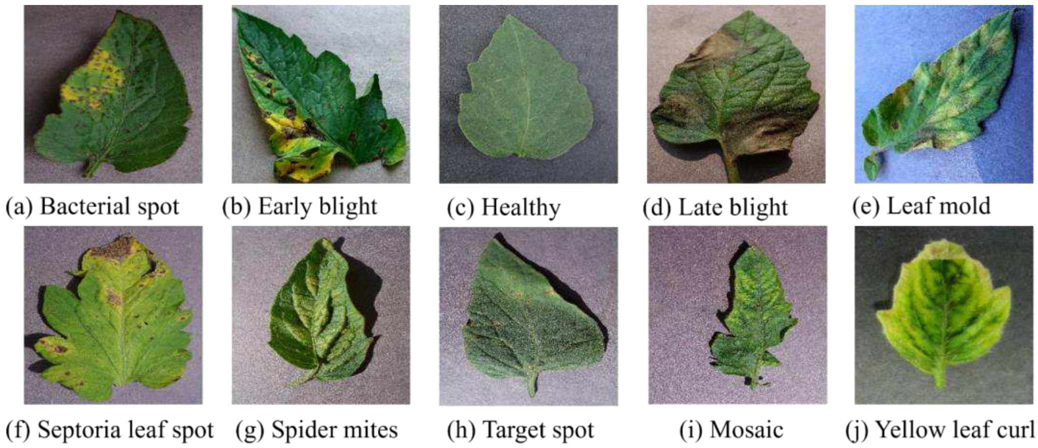



**Figure 18.** Samples of four different tomato diseases

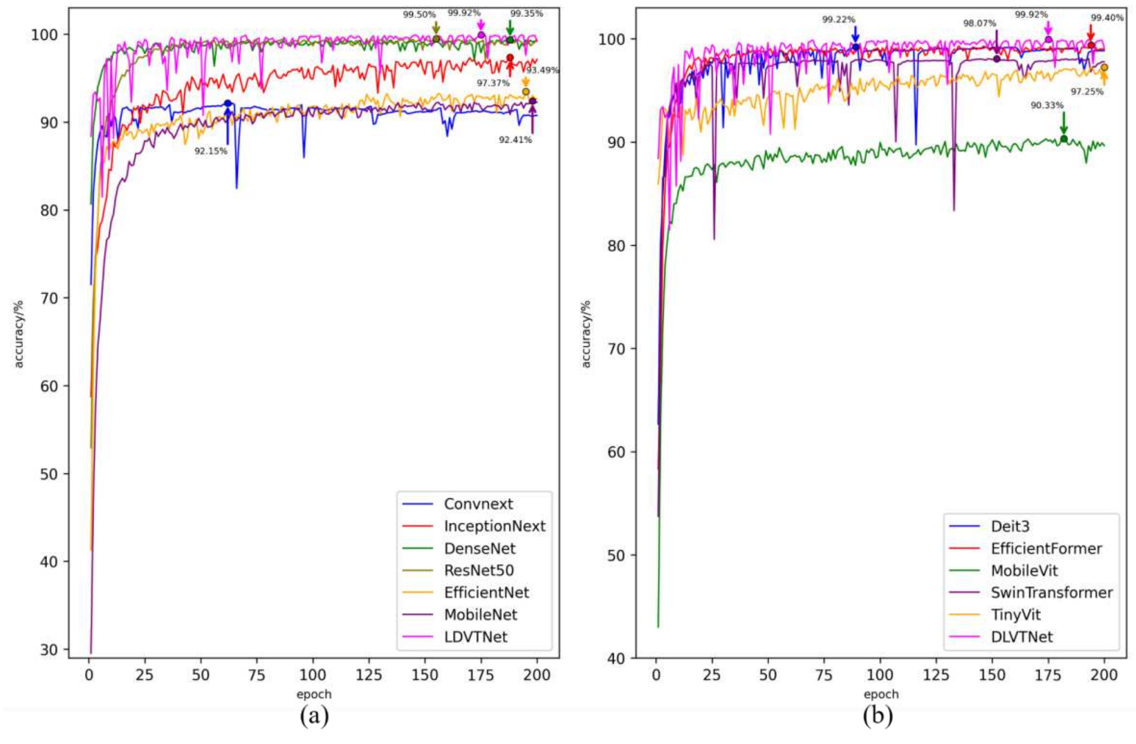



**Figure 19.** The change of training accuracy curves of different models. (a) is the comparison of CNN models and (b) is the comparison of Transformer models

## 4. Conclusions

In this paper, a deep learning-based method for grape leaf disease recognition and detection is proposed, which effectively identifies different grape leaf diseases and enhances the richness of data samples. Firstly, to address the issue that existing data augmentation methods struggle to effectively enrich image dataset information, FastGAN adversarial generative networks were used to generate a large number of grape leaf samples. This approach effectively enriches the feature representation of grape leaf diseases at different stages and addresses the problem of data imbalance. Building on this, a lightweight neural network model, DLVTNet, was proposed. This model, by combining the Transformer-based LVT Block and CNN-based MARI Block, effectively integrates both global and local information in images to recognize leaf diseases. The LVT Block innovatively introduces the Ghost Module to enhance image information extraction. Additionally, based on the low dependency of K and Q on the number of image channels in self-attention mechanisms, a Channel-wise Local Self-Attention mechanism (CLSHSA) was proposed to achieve lightweight self-attention without compromising the model's recognition capability. Furthermore, the inverted residual blocks were improved by incorporating the MELA attention mechanism, which utilizes positional and multi-scale information in images to obtain weights, thus enhancing the MARI Block's ability to extract local information. Lastly, the paper used dense connections to link LVT Block and MARI Block, increasing the reuse of image information within the DLVTNet model and further improving the model's recognition capability.

To validate the performance of the DLVTNet model, various experiments were conducted, including model comparisons, ablation studies, and experiments with different attention mechanisms. Experimental results confirm that using FastGAN adversarial generative networks effectively enriches sample information in the dataset and improves the model's clustering effect. The proposed DLVTNet network model, which combines CNN and Transformer structures, demonstrates superior recognition ability compared to mainstream image classification models, while significantly reducing computational resource requirements and model parameters. Additionally, the MELA attention mechanism proposed in this paper accurately captures disease regions in leaves compared to mainstream attention mechanisms. Finally, comparison experiments using the tomato disease dataset demonstrated that the DLVTNet model maintains high recognition accuracy and outperforms mainstream neural network models. However, there is some fluctuation in accuracy curves during the early stages of training, which requires further exploration of model structure, training methods, and dataset optimization to improve DLVTNet's recognition ability and reduce fluctuations in accuracy curves during training. Overall, the deep learning-based grape leaf disease detection method proposed in this paper effectively achieves disease detection and provides new research ideas for modern agricultural leaf disease detection and the integration of CNN and Transformer models.

**Author Contributions:** Conceptualization, Ning Zhang.; methodology, Enxu Zhang.; validation, Fei Li.; writing—review and editing, Guowei Qi.; visualization, Cheng Lv.;

**Funding:** This research received no external funding

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

**Data availability statement:** The data that support the findings of this study are available on request from the corresponding author, Enxu Zhang, upon reasonable request.

## References

1 Mira de Orduña, R. Climate change associated effects on grape and wine quality and production. *Food Research International* **43**, 1844-1855, doi:10.1016/j.foodres.2010.05.001 (2010).

2 Burr, T. J., Bazzi, C., Süle, S. & Otten, L. Crown Gall of Grape: Biology of Agrobacterium vitis and the Development of Disease Control Strategies. *Plant disease* **82 12**, 1288-1297, doi:10.1094/PDIS.1998.82.12.1288 (1998).

3 Prathibha, S. R., Hongal, A. & Jyothi, M. P. in *2017 International Conference on Recent Advances in Electronics and Communication Technology (ICRAECT).* 81-84.

4 Zhanga, Z. *et al.* Image Recognition of Maize Leaf Disease Based on GA-SVM. *Chemical engineering transactions* **46**, 199-204, doi:10.3303/CET1546034 (2015).

5 Tan, A., Zhou, G., He, M. J. M. T. & Applications. Surface defect identification of Citrus based on KF-2D-Renyi and ABC-SVM. *Multimedia Tools and Applications* **80**, 9109 - 9136, doi:10.1007/s11042-020-10036-y (2021).

6 Khan, M. A. *et al.* CCDF: Automatic system for segmentation and recognition of fruit crops diseases based on correlation coefficient and deep CNN features. *Computers and Electronics in Agriculture* **155**, 220-236, doi:10.1016/j.compag.2018.10.013 (2018).

7 Chen, X. *et al.* Identification of tomato leaf diseases based on combination of ABCK-BWTR and B-ARNet. *Computers and Electronics in Agriculture* **178**, 105730, doi:10.1016/j.compag.2020.105730 (2020).

8 Goncharov, P., Uzhinskiy, A., Ososkov, G., Nechaevskiy, A. & Zudikhina, J. Deep Siamese Networks for Plant Disease Detection. *EPJ Web of Conferences*, doi:10.1051/epjconf/202022603010 (2020).

9 Bao, W., Yang, X., Liang, D., Hu, G. & Yang, X. Lightweight convolutional neural network model for field wheat ear disease identification. *Computers and Electronics in Agriculture* **189**, 106367 (2021).

10 Atila, Ü., Uçar, M., Akyol, K. & Uçar, E. Plant leaf disease classification using EfficientNet deep learning model. *Ecological Informatics* **61**, 101182, doi:10.1016/j.ecoinf.2020.101182 (2021).

11 Lin, J. *et al.* GrapeNet: A Lightweight Convolutional Neural Network Model for Identification of Grape Leaf Diseases. *Agriculture* (2022).

12 Fang, S. *et al.* Multi-channel feature fusion networks with hard coordinate attention mechanism for maize disease identification under complex backgrounds. *Computers and Electronics in Agriculture* **203**, 107486 (2022).
13 Zhang, Y., Huang, S., Zhou, G., Hu, Y. & Li, L. Identification of tomato leaf diseases based on multi-channel automatic orientation recurrent attention network. *Computers and Electronics in Agriculture* **205**, 107605, doi:10.1016/j.compag.2022.107605 (2023).
14 Sharma, V., Tripathi, A. K. & Mittal, H. DLMC-Net: Deeper lightweight multi-class classification model for plant leaf disease detection. *Ecological Informatics* **75**, 102025, doi:10.1016/j.ecoinf.2023.102025 (2023).
15 Zhao, S., Peng, Y., Liu, J. & Wu, S. Tomato Leaf Disease Diagnosis Based on Improved Convolution Neural Network by Attention Module. *Agriculture* **11**, 651, doi:10.3390/agriculture11070651 (2021).
16 Zhao, Y., Sun, C., Xu, X. & Chen, J. RIC-Net: A plant disease classification model based on the fusion of Inception and residual structure and embedded attention mechanism. *Computers and Electronics in Agriculture* **193**, 106644, doi:10.1016/j.compag.2021.106644 (2022).
17 Zeng, W. & Li, M. Crop leaf disease recognition based on Self-Attention convolutional neural network. *Computers and Electronics in Agriculture* **172**, 105341, doi:10.1016/j.compag.2020.105341 (2020).
18 Chen, J., Zhang, D., Zeb, A. & Nanehkaran, Y. Identification of rice plant diseases using lightweight attention networks. *Expert Systems with Applications* **169**, 114514, doi:10.1016/j.eswa.2020.114514 (2021).
19 Vallabhajosyula, S., Sistla, V. & Kolli, V. K. K. A novel hierarchical framework for plant leaf disease detection using residual vision transformer. *Heliyon* **10**, e29912, doi:10.1016/j.heliyon.2024.e29912 (2024).
20 Waheed, A. *et al.* An optimized dense convolutional neural network model for disease recognition and classification in corn leaf. *Computers and Electronics in Agriculture* **175**, 105456, doi:10.1016/j.compag.2020.105456 (2020).
21 Chen, Z., Wang, G., Lv, T. & Zhang, X. Using a Hybrid Convolutional Neural Network with a Transformer Model for Tomato Leaf Disease Detection. *Agronomy* **14**, 673, doi:10.3390/agronomy14040673 (2024).
22 Han, D. & Guo, C. Automatic classification of ligneous leaf diseases via hierarchical vision transformer and transfer learning. *Frontiers in Plant Science* **14**, doi:10.3389/fpls.2023.1328952 (2024).
23 Hu, B., Jiang, W., Zeng, J., Cheng, C. & He, L. FOTCA: hybrid transformer-CNN architecture using AFNO for accurate plant leaf disease image recognition. *Frontiers in Plant Science* **14**, doi:10.3389/fpls.2023.1231903 (2023).
24 Bhaldar, H. Novel Algorithm for Detection and Classification of Grape Leaf Disease Using K-Mean Clustering. *2nd National Conference on Recent Advances in Engineering and Technology (NCRAET_17)* (2007).
25 Phookronghin, K., Srikaew, A., Attakitmongcol, K. & Kumsawat, P. 2 Level Simplified Fuzzy ARTMAP for Grape Leaf Disease System Using Color Imagery and Gray Level Co-Occurrence Matrix. *2018 International Electrical Engineering Congress (iEECON)*, 1-4, doi:10.1109/IEECON.2018.8712183 (2018).
26 Mohammed, K. K., Darwish, A. A. & Hassenian, A. E. Artificial Intelligent System for Grape Leaf Diseases Classification. *Computer Science, Agricultural and Food Sciences*, 19–29, doi:10.1007/978-3-030-51920-9_2 (2020).
27 Ji, M., Zhang, L. & Wu, Q. Automatic grape leaf diseases identification via UnitedModel based on multiple convolutional neural networks. *Information Processing in Agriculture* **7**, 418-426, doi:10.1016/j.inpa.2019.10.003 (2020).
28 Suo, J. *et al.* CASM-AMFMNet: A Network Based on Coordinate Attention Shuffle Mechanism and Asymmetric Multi-Scale Fusion Module for Classification of Grape Leaf Diseases. *Frontiers in Plant Science* **13**, doi:10.3389/fpls.2022.846767 (2022).
29 Liu, B. *et al.* Grape Leaf Disease Identification Using Improved Deep Convolutional Neural Networks. *Frontiers in Plant Science* **11**, doi:10.3389/fpls.2020.01082 (2020).
30 Cai, C. *et al.* Identification of grape leaf diseases based on VN-BWT and Siamese DWOAM-DRNet. *Engineering Applications of Artificial Intelligence* **123**, 106341, doi:10.1016/j.engappai.2023.106341 (2023).

31 Karthik, R. *et al.* GrapeLeafNet: A Dual-Track Feature Fusion Network With Inception-ResNet and Shuffle-Transformer for Accurate Grape Leaf Disease Identification. *IEEE Access* **12**, 19612-19624, doi:10.1109/ACCESS.2024.3361044 (2024).

32 Zeng, W., Li, H., Hu, G. & Liang, D. Lightweight dense-scale network (LDSNet) for corn leaf disease identification. *Computers and Electronics in Agriculture* **197**, 106943, doi:10.1016/j.compag.2022.106943 (2022).

33 Yun, S. & Ro, Y. SHViT: Single-Head Vision Transformer with Memory Efficient Macro Design. *ArXiv* **abs/2401.16456**, doi:10.48550/arXiv.2401.16456 (2024).

34 Cai, H., Li, J., Hu, M., Gan, C. & Han, S. EfficientViT: Lightweight Multi-Scale Attention for High-Resolution Dense Prediction. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 17256-17267, doi:10.1109/ICCV51070.2023.01587 (2023).

35 Hou, Q., Daquan, Z. & Feng, J. *Coordinate Attention for Efficient Mobile Network Design*. (2021).

36 Woo, S. *et al.* in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 16133-16142.

37 Tan, M. & Le, Q. V. EfficientNetV2: Smaller Models and Faster Training. *International Conference on Machine Learning(ICML 2021)*, doi:10.48550/arXiv.2104.00298 (2021).

38 Howard, A. *et al.* Searching for MobileNetV3. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1314-1324, doi:10.1109/ICCV.2019.00140 (2019).

39 He, K., Zhang, X., Ren, S. & Sun, J. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 770-778.

40 Huang, G., Liu, Z., Maaten, L. V. D. & Weinberger, K. Q. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2261-2269.

41 Yu, W., Zhou, P., Yan, S. & Wang, X. *InceptionNeXt: When Inception Meets ConvNeXt*. (2023).

42 Touvron, H., Cord, M. & J'egou, H. e. in *European Conference on Computer Vision.*

43 Li, Y. *et al.* in *2023 IEEE/CVF International Conference on Computer Vision (ICCV).* 16843-16854.

44 Mehta, S. & Rastegari, M. Separable Self-attention for Mobile Vision Transformers. *ArXiv* **abs/2206.02680**, doi:10.48550/arXiv.2206.02680 (2022).

45 Liu, Z. *et al.* in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 11999-12009.

46 Wu, K. *et al.* in *Computer Vision – ECCV 2022.* (eds Shai Avidan *et al.*) 68-85 (Springer Nature Switzerland).

47 Wang, Q. *et al.* in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 11531-11539.

48 Woo, S., Park, J., Lee, J.-Y. & Kweon, I.-S. CBAM: Convolutional Block Attention Module. *ArXiv* **abs/1807.06521**, doi:10.48550/arXiv.1807.06521 (2018).

49 Liu, Y., Shao, Z. & Hoffmann, N. Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions. *ArXiv* **arXiv:2112.05561** doi:10.48550/arXiv.2112.05561 (2021).

50 Liu, Y., Shao, Z., Teng, Y. & Hoffmann, N. NAM: Normalization-based Attention Module. *ArXiv* **abs/2111.12419**, doi:10.48550/arXiv.2111.12419 (2021).

51 Xu, W. & Wan, Y. ELA: Efficient Local Attention for Deep Convolutional Neural Networks. *ArXiv* **arXiv:2403.01123**, doi:10.48550/arXiv.2403.01123 (2024).