frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 28 April 2022
doi: 10.3389/fpls.2022.864486

Deep Learning-Based Identification of Maize Leaf Diseases Is Improved by an Attention Mechanism: Self-Attention

*Xiufeng Qian[1,2,3], Chengqi Zhang[4], Li Chen[4] and Ke Li[1,2,3]**

[1] School of Information and Computer, Anhui Agricultural University, Hefei, China, [2] Anhui Provincial Engineering Laboratory for Beidou Precision Agriculture Information, Anhui Agricultural University, Hefei, China, [3] Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui University, Hefei, China, [4] School of Plant Protection, Anhui Agricultural University, Hefei, China

Maize leaf diseases significantly reduce maize yield; therefore, monitoring and identifying the diseases during the growing season are crucial. Some of the current studies are based on images with simple backgrounds, and the realistic field settings are full of background noise, making this task challenging. We collected low-cost red, green, and blue (RGB) images from our experimental fields and public dataset, and they contain a total of four categories, namely, southern corn leaf blight (SCLB), gray leaf spot (GLS), southern corn rust (SR), and healthy (H). This article proposes a model different from convolutional neural networks (CNNs) based on transformer and self-attention. It represents visual information of local regions of images by tokens, calculates the correlation (called attention) of information between local regions with an attention mechanism, and finally integrates global information to make the classification. The results show that our model achieves the best performance compared to five mainstream CNNs at a meager computational cost, and the attention mechanism plays an extremely important role. The disease lesions information was effectively emphasized, and the background noise was suppressed. The proposed model is more suitable for fine-grained maize leaf disease identification in a complex background, and we demonstrated this idea from three perspectives, namely, theoretical, experimental, and visualization.

Keywords: crop disease, machine learning, deep learning, attention mechanism, neural network

OPEN ACCESS

Edited by:
Valerio Giuffrida,
Edinburgh Napier University,
United Kingdom

Reviewed by:
Marcin Wozniak,
Silesian University of Technology,
Poland
Huaming Chen,
The University of Adelaide, Australia

***Correspondence:**
Ke Li
like@ahau.edu.cn

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

Received: *28 January 2022*
Accepted: *28 March 2022*
Published: *28 April 2022*

Citation:
Qian X, Zhang C, Chen L and Li K (2022) Deep Learning-Based Identification of Maize Leaf Diseases Is Improved by an Attention Mechanism: Self-Attention. Front. Plant Sci. 13:864486. doi: 10.3389/fpls.2022.864486

INTRODUCTION

Maize is one of the most important crops for humanity, with the highest yield globally (Ranum et al., 2014). Maize diseases can cause severe yield reductions, a critical problem (Savary et al., 2012). Therefore, it is vital to promptly identify and monitor maize diseases during the growing period. Accurate identification of diseases in maize is difficult for crop growers who may not be professional in plant pathology, and expert identification is expensive and time-consuming (Ouppaphan, 2017). Traditional image recognition methods and deep learning are gradually entering the field of plant disease recognition (Saleem et al., 2019).

Mobile terminals based on web services and support vector machine (SVM) as back-end algorithms can automatically identify maize diseases (Zhang and Yang, 2014). Zhang et al. (2015) proposed an improved genetic algorithm-SVM (GA-SVM) algorithm to improve

the accuracy. A recent study on maize disease identification compared five standard machine learning methods (Panigrahi et al., 2020), namely, Naive Bayes (NB), Decision Tree (DT), K-Nearest Neighbor (KNN), SVM, and Random Forest (RF), with RF achieving the highest accuracy of 79.23%.

However, traditional machine learning is mainly limited by feature extraction and feature representation. Deep learning has made significant progress in plant disease identification (Liu and Wang, 2021). Since AlexNet was proposed in 2012 (Krizhevsky et al., 2012), convolutional neural networks (CNNs) have been widely used in academia and industry, e.g., face detection in dangerous situations (Wieczorek et al., 2021) and combination of Internet of things (IoT) and pearl millet disease prediction (Kundu et al., 2021). In the field of plant disease identification, Dhaka et al. (2021) provided a systematic review of relevant deep learning techniques. Due to its low complexity, a lightweight CNN for mobile terminals has achieved satisfactory performance in maize disease identification (Ouppaphan, 2017). A CNN-based system (DeChant et al., 2017) was implemented to automatically identify northern leaf blight, addressing the challenges of limited data and various irregularities appearing in field-grown images. Ahila Priyadharshini et al. (2019) proposed a CNN modified from LeNet for identifying four maize categories (three diseases classes and one health class) with an accuracy of 97.89%.

However, most of the current studies are based on simple background maize leaf or other crop disease recognition, and the recognition effect of the trained models deteriorates in real field settings, because background noise information causes serious obstruction (Lv et al., 2020). Current research on popular or novel deep learning image recognition algorithms (CNNs) is mainly tested on the public dataset ImageNet, and its images are different from fine-grained images of crop disease. Those designed CNNs mostly focus on patterns of objects in images (e.g., profile features of dogs or cats), and these pattern features are reflected in feature maps of convolutional output, as can be demonstrated by numerous neural network visualization studies (Chattopadhay et al., 2018; Chen et al., 2020; Jiang et al., 2021). In contrast, fine-grained crop disease lesions are usually similar and discrete on the leaf surface; thus, CNNs may not be fully adapted to fine-grained maize leaf disease image classification tasks, which will result in no increase in model performance even by stacking the network layers and increasing model parameters. Rational model design for specific tasks is important and necessary, and the following analysis and experiments in this article also prove this perspective. In addition, many visual disturbances (e.g., reflection, dispersion, and blur) seriously affect fine-grained image classification (Lu Y. et al., 2017; Yang et al., 2020). Therefore, fine-grained maize disease identification in complex background field settings requires more rational models and computerized mechanisms.

Mutual attention between words is highly essential for machine translation tasks, which determines whether a sentence can be translated accurately. The transformer architecture (Vaswani et al., 2017) with the attention mechanism has achieved significant success in natural language processing (NLP). Although previous attention mechanisms have been applied to some specific tasks, e.g., image caption generation technology (Lu J. et al., 2017), text classification (Li et al., 2019), and human action recognition (Song et al., 2017), the form and principle of their attention mechanisms are too different and specialized. However, the transformer's attention mechanism (self-attention) has a universal form.

To explore whether the attention mechanism will bring enhancements to the field of computer vision, vision transformer (ViT, **Figure 1** depicts it) (Dosovitskiy et al., 2020) applies the transformer architecture directly to image classification tasks for the first time, outperforming the state of the art on large-scale datasets. Subsequently, researchers gradually began to study ViT and its attention mechanism. Transformer in transformer (TNT) (Han et al., 2021) embeds the inner transformer into the outer transformer to improve the feature extraction capability lacking in the patch embedding method (refer to **Figure 1** for the patch embedding method). Compact convolutional transformer (CCT) (Hassani et al., 2021) demonstrates that convolution can be used to extract local information better, thus making it possible to apply transformer to more tasks with small datasets.

In this study, we found that transformer and self-attention computer mechanisms are more suitable for maize leaf disease identification in complex backgrounds. This article will demonstrate their efficiency and why they work from three perspectives, namely, theoretical derivation, experiment, and visualization. We collected maize leaf diseases datasets with complex backgrounds in our experimental field and proposed an improved model based on ViT and CCT to classify maize into four categories (**Figure 2**), namely, healthy (H), southern corn leaf blight (SCLB) (Aregbesola et al., 2020), gray leaf spot (GLS) (Saito et al., 2018), and southern corn rust (SR) (Wang S. et al., 2019). The model outperforms some mainstream CNNs compared with it in all metrics, with a smaller number of parameters. In addition, we also conducted experiments on the necessity of the self-attention for the model, demonstrating that it is essential. This article also conducts experiments to observe the effect of the ratio of train set to validation set on the accuracy of the model.

The rest of this article is organized as follows. The section "Materials and Methods" introduces the details of our experimental field and experimental sample cultivation, describing our datasets and methods used to collect them. In that section, we focused on describing our algorithm and the detailed theoretical derivation and proving its effectiveness, as well as the experimental visualization schemes (three schemes). All experimental results are described in section "Results." We discussed the reasons for the efficiency of this model and some possible future extensions in section "Discussion."

MATERIALS AND METHODS

Data Collection and Preparation

The dataset of the images, which included 7,701 images, consists of two parts, namely, one part is collected from the public dataset Plant Village and the other part is taken by mobile phones in the natural environment of our experimental field. The maize plants grown in the experimental field are used to select suitable

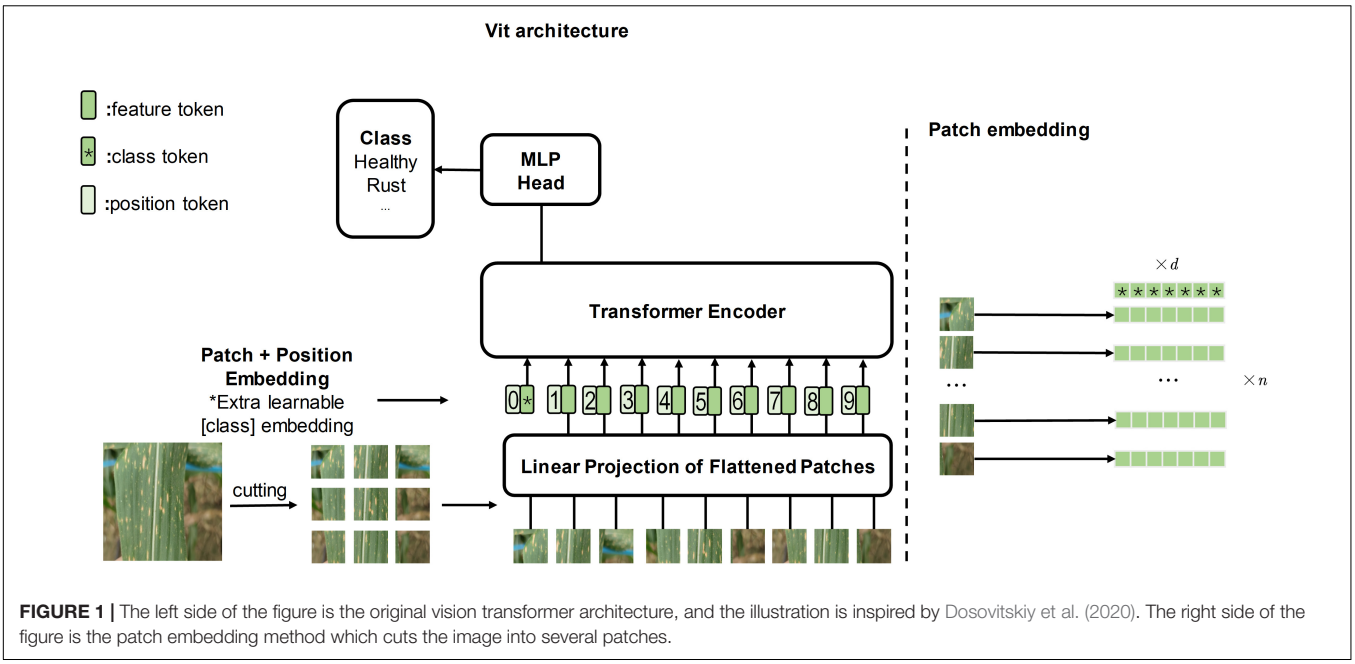

FIGURE 1 | The left side of the figure is the original vision transformer architecture, and the illustration is inspired by Dosovitskiy et al. (2020). The right side of the figure is the patch embedding method which cuts the image into several patches.

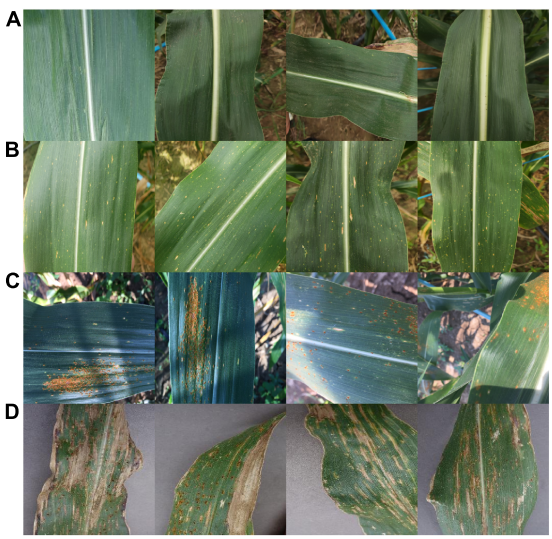

FIGURE 2 | Showing four categories of samples of maize. **(A)** H. **(B)** SCLB. **(C)** SR. **(D)** GLS.

disease-resistant varieties, so there are numerous maize varieties. However, as with other studies on maize disease identification, the variety of maize is not the focus and has no impact on the study of this article because the images of maize leaves in our dataset do not reflect their genetic variety. An area of the experimental field covered 3 acres was chosen for this study, planting a total of 80 rows of maize with 26 maize plants per row, 65 cm between rows, and 13 m length of each row. Half of this area was planted with maize inoculated with SCLB, and the other half with maize inoculated with SR. The conidia with a concentration of 10^6/ml were sprayed at this maize in the sixth-leaf stage, namely, 40–50 days after sowing, to inoculate maize with the abovementioned diseases. After inoculation, the maize is allowed to develop naturally. One day of the milk stage of the maize is chosen to take all the images needed for our dataset. Every maize plant is sampling points. We walked along the rows and remained for several minutes to take images, and the same leaf will be photographed more than once to get 1–6 images. Furthermore, the leaves were manually moved to find a better angle to photograph a good image while adjusting the position of the phones to aid this operation. Despite the fact that a leaf may be photographed more than once, every image is different and contains complicated background visual information because the content of interest is different for each shot. The manual focus is chosen to solve the issue that phones cannot focus on the leaf lesion areas of interest, therefore, guaranteeing every image is clear and focused. The H maize images were obtained from another area of the experimental field where eight rows of maize plants were planted, and the planting pattern and the photographing mode are identical to the above. All the images photographed are under normal uncontrolled lighting conditions with mobile phones' low-cost red, green, and blue (RGB) sensor. The GLS maize images are downloaded from the Plant Village. This article divided the dataset into a training dataset and a validation dataset according to the principles of 3 to 1 due to the sample balance. **Table 1** shows the distribution of images and the division of the dataset.

Data Processing

The images' size must be unified to a standard 224 × 224-pixel square offline to reduce the computational effort before the model training. Furthermore, some data augmentation techniques are separately applied to each image, with a certain probability

TABLE 1 | Distribution of data sources and division of training set and validation set.

Categories	Shooting by us	Plant village	Train set	Validation set
SCLB	2,243	0	1,743	500
H	1,273	1,162	1,953	500
SR	2,023	0	1,523	500
GLS	0	1,000	750	250

during model training online, thus enhancing the generalization ability of the model and preventing its overfitting. This article selects four data augmentation techniques suitable for maize leaf disease identification, namely, RandomFlip, ColorJitter, Cutmix (Yun et al., 2019), and Mixup (Zhang et al., 2018). Before an image is imported into the model, RandomFlip randomly rotates it horizontally or vertically, expanding the dataset, as this is equivalent to the images in the dataset having different shooting angles than their raw form. ColorJitter randomly changes the image's brightness, contrast, saturation, and hue. As a result, ColorJitter can improve the model's ability to adapt to different lights in field settings. The lesions of the three diseases chosen for the study are scattered on the surface of the leaves, which means that the model should not focus only on the lesions of one area but also on the entire leaf. Cutmix randomly crops a patch of the image and fills the area with a small and same size patch from another image. The size of the patches is a hyperparameter, and the position of the patches on the images is random. Mixup is widely used in image classification tasks, and it mainly constructs a virtual sample $(\widetilde{x}, \widetilde{y})$ by the following methods:

$$\widetilde{x} = \lambda x_i + (1-\lambda)\, x_j \tag{1}$$

$$\widetilde{y} = \lambda y_i + (1-\lambda)\, y_j \tag{2}$$

where x_i and x_j are two different images, y_i and y_j are the unique one-hot labels corresponding to these two images, and $\lambda \in (0, 1)$. Mixup extends the distribution of samples by linear interpolation, making it popular for various image classification tasks. Both Cutmix and Mixup make models confusing, forcing them to focus on global information rather than local information. This article's data augmentation techniques are used only in training, not testing.

Algorithm

At the beginning of this section, we have done some specifications of mathematical notation and some pre-paving for our model. The upper case non-bolded symbols in this article refer to matrices, the lower case bolded symbols refer to row vectors, and the lower case non-bolded symbols refer to constants or scalar variables. A complete image can be divided into several local regions. The critical feature information of maize leaf disease is located in some local regions where the lesions are located. From the visual point of view, the texture and color of these local areas are the feature information. From the algorithmic point of view, the RGB values of the pixels in these local areas are the feature information. Background information that interferes with the classification is useless information. CNNs usually downsample the image and use the generated feature maps to represent the information of the image. Our model encodes the feature information of local regions into vectors (called tokens) to represent the information of the whole image. The attention mechanism of this article will be based on these tokens to identify those critical regions to make the classification.

Our standard model has three stages, namely, Stage 1, Stage 2, and Stage 3 (**Figure 3**). Stage 1 extracts the image features and encodes them into a feature tokens matrix. Each row vector in the tokens matrix is a token, and a token is a vector used to represent the local visual features within a receptive field (convolution or max-pooling kernel). Passing the input image $I \in R^{h \times w \times c}$ through a convolutional layer and a max-pooling layer generates feature maps $Fm \in \mathbb{R}^{l \times l \times d}$ with channels of d. The width of feature maps output from both convolution layer and max-pooling layer is expressed by the following equation:

$$l = \frac{i+2p-k}{s}+1 \tag{3}$$

where i denotes the width of the original image or input feature maps, k is the size of the kernel (convolution or max-pooling), s is the stride of kernel movement, and p is padding. We listed those hyperparameters at the end of the section algorithm. At the end of Stage 1, after extracting vectors along the channel dimension for the feature maps Fm, the vectors are arranged to obtain the feature tokens matrix, which can be described by the following equation:

$$X = Flatten\,(Fm) \tag{4}$$

where $X \in \mathbb{R}^{n \times d}$ is the tokens matrix, and $n = l^2$. Each row vector of dimension d in X is a feature token.

At the beginning of Stage 2, a learnable vector "classification token" of dimension d is appended to the top of X; hence, $X \in \mathbb{R}^{n \times d}$ was transformed to $X \in \mathbb{R}^{n_t \times d}$, where $n_t = n + 1$. The "classification token" is derived from NLP and is similar to BERT's (Devlin et al., 2018) "class token." The classification token will be output at the end of Stage 2 as input to Stage 3 to complete the final classification. Therefore, the transformer encoder of Stage 2 is the core computational module of the whole network, and the essential part of it is multi-head self-attention (MSA) that is used to perform self-attention. The rest of the section algorithm will introduce how it works, demonstrating why it is effective. To better explain MSA, we first described the computational process of single-head self-attention (SSA). Tokens matrix X is linearly transformed into queries Q, keys K, and values V by three matrices, W_Q, W_K, and W_V, respectively, and the linear transforms can be seen in the following equations:

$$Q = XW_Q \tag{5}$$

$$K = XW_K \tag{6}$$

$$V = XW_V \tag{7}$$

where $W_Q \in \mathbb{R}^{d \times d}$, $W_K \in \mathbb{R}^{d \times d}$, and $W_V \in \mathbb{R}^{d \times d}$ are parametric learnable matrices. In fact, each row vector in Q, K,

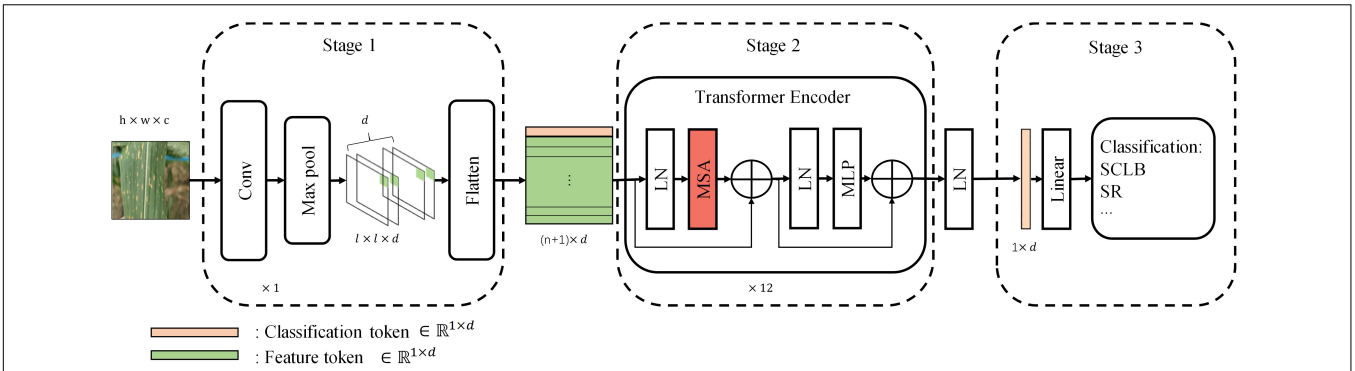

FIGURE 3 | The standard model architecture. Stage 1 extracts information from local regions of the image and encodes them into tokens. Stage 2 is the core computational network that performs the self-attention. Stage 3 maps the classification token into four classes to complete the final classification. Linear, linear layer; LN, layer normalization; MLP, multilayer perceptron.

and V is still a token used to represent feature information of the corresponding local region. Assume that $\mathbf{q_i}$, $\mathbf{k_i}$, and $\mathbf{v_i}$ denote the i-th token of Q, K, and V, respectively; thus, they all represent the feature information of the i-th receptive field of the original image. The correlation between tokens is obtained by calculating the inner product of all row vectors in Q and all row vectors in K. For example, $\langle \mathbf{q_i}, \mathbf{k_j} \rangle = \mathbf{q_i}\mathbf{k}_j^T$ represents the correlation between the i-th token and the j-th token or the degree of attention of the i-th token to the j-th token. However, it is usually not equal to $\langle \mathbf{q_j}, \mathbf{k_i} \rangle = \mathbf{q_j}\mathbf{k}_i^T$, which is due to two factors. On the one hand, Q and K are obtained by a linear transformation of two different learnable matrices, W_Q and W_K. Although both $\mathbf{q_i}$ and $\mathbf{k_i}$ represent the visual information of the i-th receptive field, the elements in W_Q and W_K change in the direction favorable to the final classification as the model weights are updated. On the other hand, the self-attention mechanism is derived from NLP, where words are encoded as vectors (tokens) in a machine translation task. The correct translation of a sentence requires finding the relevance of each word, and two words have different attention to each other, which requires the correlation calculation method between tokens as described earlier. Therefore, the correlation between tokens can be calculated by the following equations:

$$A = QK^T$$

$$= \begin{bmatrix} \mathbf{q}_1\mathbf{k}_1^T & \mathbf{q}_1\mathbf{k}_2^T & \cdots & \mathbf{q}_1\mathbf{k}_{n_t}^T \\ \mathbf{q}_2\mathbf{k}_1^T & \mathbf{q}_2\mathbf{k}_2^T & \cdots & \mathbf{q}_2\mathbf{k}_{n_t}^T \\ \vdots & \vdots & & \vdots \\ \mathbf{q}_{n_t}\mathbf{k}_1^T & \mathbf{q}_{n_t}\mathbf{k}_2^T & \cdots & \mathbf{q}_{n_t}\mathbf{k}_{n_t}^T \end{bmatrix}$$

$$= \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n_t} \\ a_{21} & a_{22} & \cdots & a_{2n_t} \\ \vdots & \vdots & & \vdots \\ a_{n_t1} & a_{n_t2} & \cdots & a_{n_tn_t} \end{bmatrix} \quad (8)$$

A is the preliminary tokens correlation matrix; in other words, it represents the attention between tokens, e.g., a_{ij} denotes the attention of the i-th token to the j-th token or the attention of the visual information of the i-th receptive field to the visual information of the j-th receptive field. The following equations normalize the attention matrix A:

$$A' = soft\max\left(\frac{A}{\sqrt{d_k}}\right)$$

$$= \begin{bmatrix} \sigma(\mathbf{a}_1)/\sqrt{d_k} \\ \sigma(\mathbf{a}_2)/\sqrt{d_k} \\ \vdots \\ \sigma(\mathbf{a}_{n_t})/\sqrt{d_k} \end{bmatrix} \quad (9)$$

$$\sigma(\mathbf{a}_i) = \left[\frac{e^{a_{i1}}}{\sum_{j=1}^{n_t} e^{a_{ij}}} \quad \frac{e^{a_{i2}}}{\sum_{j=1}^{n_t} e^{a_{ij}}} \quad \cdots \quad \frac{e^{a_{in_t}}}{\sum_{j=1}^{n_t} e^{a_{ij}}} \right] \quad (10)$$

where d_k is a normalization factor and a hyperparameter. Assume that α_{ij} is the element in row i and column j of A'. Subsequently, elements in attention matrix A' are used as weights to linearly combine the tokens of V, which will integrate the information of the tokens they are focused on for each token. The following equation describes this process:

$$V' = A'V$$

$$= \begin{bmatrix} \alpha_{11}\mathbf{v}_1 + \alpha_{12}\mathbf{v}_2 + \cdots + \alpha_{1n_t}\mathbf{v_{n_t}} \\ \alpha_{21}\mathbf{v}_1 + \alpha_{22}\mathbf{v}_2 + \cdots + \alpha_{2n_t}\mathbf{v_{n_t}} \\ \vdots \\ \alpha_{n_t1}\mathbf{v}_1 + \alpha_{n_t2}\mathbf{v}_2 + \cdots + \alpha_{n_tn_t}\mathbf{v_{n_t}} \end{bmatrix} \quad (11)$$

Thus, the new tokens of V' are integrated with the information they pay attention to. The above describes the computation of the attention mechanism. In this process, the classification token is fully involved in the computation of the self-attention mechanism, continuously integrating information about receptive fields in a different-attention way, and finally being output for final classification. The mode using tokens to represent receptive field information and integrating tokens information is more suitable for maize leaf disease identification,

because the main characteristic of maize leaf diseases is lesions, which are usually small and widely distributed on the leaf surface. Hence, similarity exists between lesions in terms of texture and color, which is reflected in the RGB values of images. The visual information in receptive fields where lesions exist is similar, and vectors encoded are also similar, so critical information of images can be highlighted by the computational model presented earlier. Subsequent experiments and visualizations in this article will also demonstrate that the model will focus on lesions rather than background noise information. MSA is a simple extension of SSA, performing *head* SSA calculations independently of each other in parallel (**Figure 4**), and *head* is a hyperparameter. Based on the SSA presented earlier, the MSA is briefly described by the following equations:

$$Q_i = XW_i^Q \tag{12}$$

$$K_i = XW_i^K \tag{13}$$

$$V_i = XW_i^V \tag{14}$$

$$A_i^{'} = softmax\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) \tag{15}$$

$$V_i^{'} = A_i^{'} V_i \tag{16}$$

$$V^{'} = Concat\left(V_1^{'}, V_2^{'}, \cdots, V_{head}^{'}\right) = V_1^{'} \oplus V_2^{'} \oplus \cdots \oplus V_{head}^{'} \tag{17}$$

where $i = 1, 2, \cdots, head$, $W_i^Q \in \mathbb{R}^{d \times \frac{d}{head}}$, $W_i^K \in \mathbb{R}^{d \times \frac{d}{head}}$, $W_i^V \in \mathbb{R}^{d \times \frac{d}{head}}$, and $\bigoplus$ is the concatenated operation to matrices. Therefore, tokens matrix $X \in \mathbb{R}^{n_t \times d}$ is calculated by the MSA and outputs $V^{'} \in R^{n_t \times d}$.

Layer normalization (LN) (Ba et al., 2016) normalizes input tokens to speed up the convergence by the following equations:

$$LN\left(\mathbf{y_i}, \alpha, \beta\right) = \frac{\mathbf{y_i} - \mu}{\sigma} \odot \alpha + \beta \in R^{n \times d} \tag{18}$$

$$\mu = \frac{1}{d}\sum_{j=1}^{d} y_i^j \tag{19}$$

$$\sigma = \sqrt{\frac{1}{d}\sum_{j=1}^{d}\left(y_i^j - u\right)^2} \tag{20}$$

where y_i is the *i*-th token, and y_i^j refers to the *j*-th element of the *i*-th token. α and β are learnable gains and bias, respectively.

Linear layer can perform a linear transformation of the input matrix, which is described by the following equation:

$$M_o = MW + \mathbf{b} \tag{21}$$

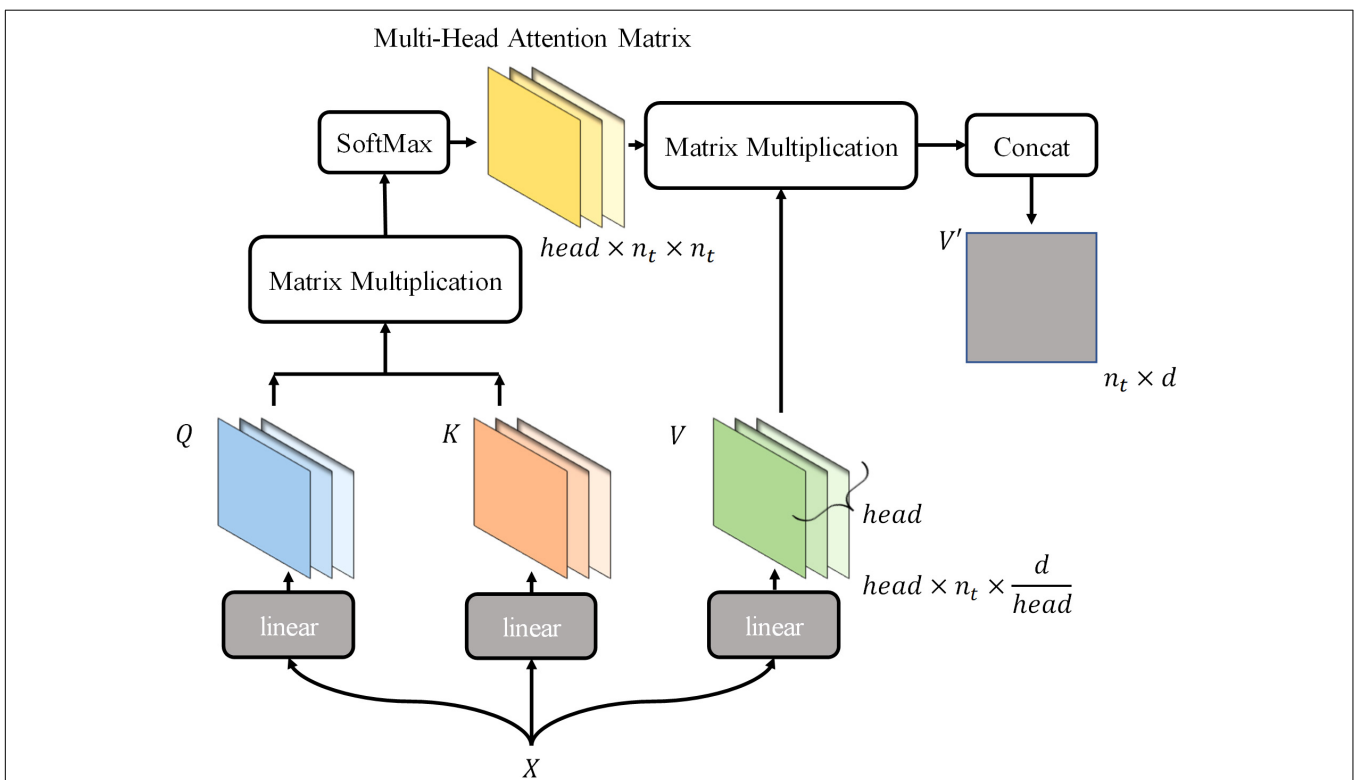

FIGURE 4 | The schematic of implementing multi-head self-attention.

where $M \in \mathbb{R}^{m \times n}$ is the input matrix, $W \in \mathbb{R}^{n \times o}$ refers to the learnable matrix, $\mathbf{b} \in \mathbb{R}^{1 \times o}$ refers to the learnable bias vector, and $M_o \in \mathbb{R}^{m \times o}$ refers to the output matrix.

Multilayer perceptron (MLP) obtains nonlinearity and transformation (Han et al., 2021), benefiting from the linear layer and the activation function Gaussian error linear units (GELU) (Hendrycks and Gimpel, 2016). This nonlinear transformation can be described as follows:

$$M_o = GELU(MW_1 + \mathbf{b}_1)W_2 + \mathbf{b}_2 \quad (22)$$

$$GELU(x) = 0.5x\left(1 + \tanh\left[\sqrt{2/\pi}(x + 0.044715x^3)\right]\right) \quad (23)$$

where $M \in \mathbb{R}^{m \times n}$ and $M \in \mathbb{R}^{m \times o}$ refer to input matrix and output matrix, respectively, $W_1 \in \mathbb{R}^{n \times h}$ and $W_2 \in \mathbb{R}^{h \times o}$ are learnable matrices, and $\mathbf{b}_1 \in \mathbb{R}^{1 \times h}$ and $\mathbf{b}_2 \in \mathbb{R}^{1 \times o}$ are learnable bias vectors. The GELU function, when applied to a matrix, will perform a nonlinear transformation on all elements of that matrix.

Stage 3 maps classification token ($\mathbf{v}_0^{'}$, the first row vector of the matrix $V^{'}$ of the last transformer encoder block) of the output of Stage 2 to four categories by a linear layer.

We conducted an ablation experiment on MSA to investigate the necessity of self-attention. The experiment needs to remove the MSA from transformer encoder. However, without the MSA, the classification tokens cannot participate in the computation of the integrated tokens information. Therefore, we designed Model-1 and Model-2 based on the standard model. Model-1 does not use classification tokens but integrates feature tokens to classify, and Model-2 removes MSA from Model-1 (refer to **Figure 5** for Model-1 and Model-2).

Hyperparameters and Training Facilities

The hyperparameters of our standard model are as follows:

- Max pool layer: number = 1, kernel size = 3, stride = 2, padding = 1,
- Convolutional layer: number = 1, kernel size = 7, stride = 4, padding = 1,
- Transformer encoder blocks: 12,
- Heads of MSA: 4,
- Dimension of token: d 64,
- Normalization factor: d_k = 16,
- Batch size: 64,
- Learning rate: 0.004,
- Weight decay: 0.05.

We have made our dataset and code, as well as all the trained models of this article, publicly available in the site: https://github.com/haiyang-qian/code-and-dataset. Our model is trained on the open-source deep learning framework Pytorch 1.9, and the programming language is Python 3.7.10. Our experimental facilities are as follows:

- CPU: Xeon Gold 6142
- GPU: RTX 3090
- CUDA: V11.2
- OS: Ubuntu 20.04
- Memory: 60.9 GB
- SSD: 429.5 GB

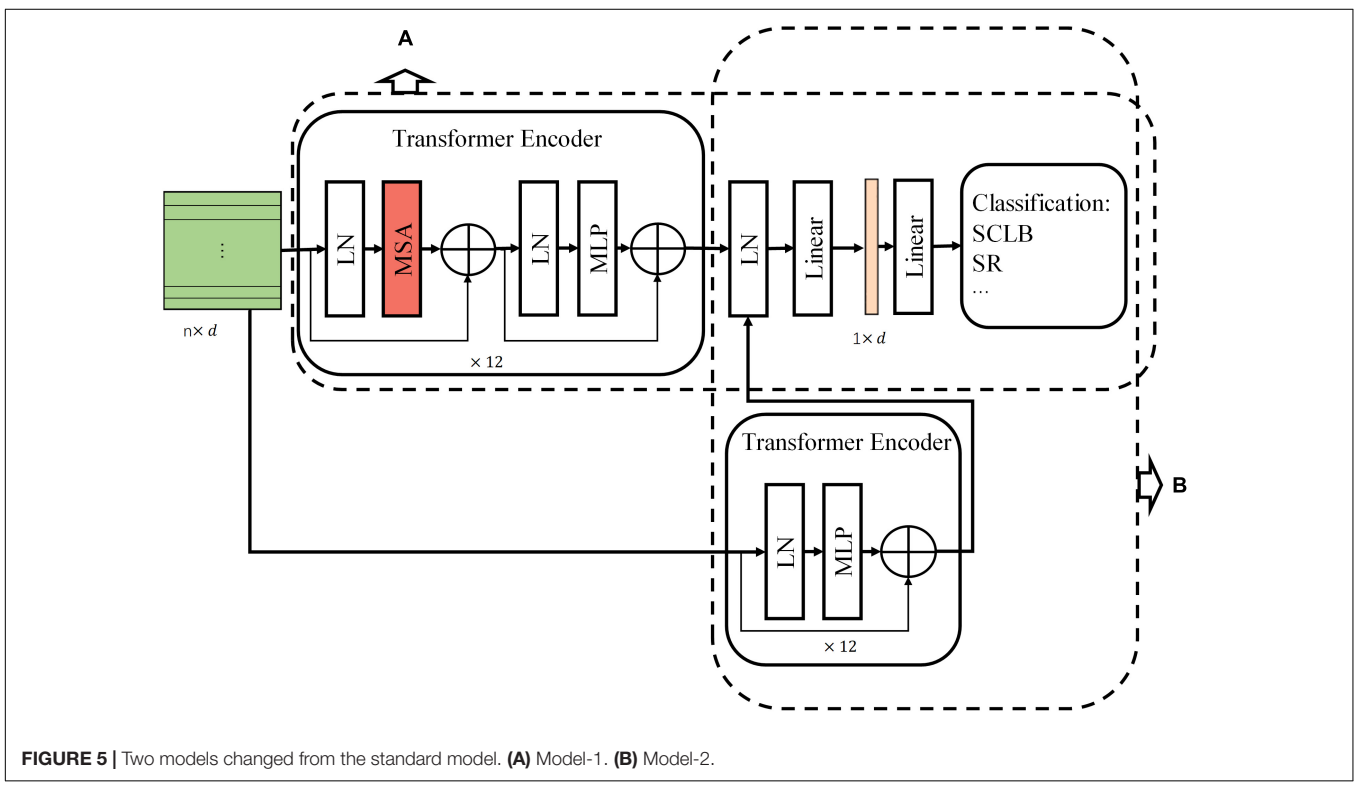

FIGURE 5 | Two models changed from the standard model. **(A)** Model-1. **(B)** Model-2.

Visualization Methods

In this article, three visualization schemes are designed, targeting three network outputs, namely, convolutional or pooling layers, tokens matrix, and classification token for feature tokens' attention. First, for the feature maps of the output of the convolution or pooling layer, we have applied the Grad-Cam method (Selvaraju et al., 2017). The method first computes gradients for class c regarding feature maps Fm of a convolutional layer (assume that Fm^k is the k-th channel of the feature maps). These gradients are globally averaged over the corresponding channels of Fm to obtain the weights of that channel α_k^c. α_k^c is the importance of feature map Fm^k for class c and is used to weigh the feature map Fm^k. Then, the class discriminative localization map (CDLM) (a map of the importance of different regions of input image for class c) can be obtained by completing this operation for all the feature maps. The above computation can be described by the following equations:

$$\alpha_k^c = \frac{1}{z}\sum_i\sum_j \frac{y^c}{Fm_{ij}^k} \tag{24}$$

$$L_{Grad-CAM}^c = \mathrm{Re}LU\left(\sum_k \alpha_k^c Fm^k\right) \tag{25}$$

$$\mathrm{Re}LU(x) = \begin{cases} x \; if \; x > 0 \\ 0 \; if \; x < 0 \end{cases} \tag{26}$$

where $L_{Grad-CAM}^c$ is a CDLM calculated by the Grad-Cam method, and it will be mapped back to the input image to obtain the visualization result. The Grad-Cam method is usually used for feature maps of the convolution or pooling layer output. Therefore, in our second visualization scheme based on the tokens matrix, we have reshaped the two-dimensional feature tokens matrix Y into a three-dimensional feature map matrix, expressed as the following mapping:

$$Y \in \mathbb{R}^{n \times d} \rightarrow Fm \in \mathbb{R}^{l \times l \times d} \tag{27}$$

where $n = l^2$. We applied the Grad-Cam method to Fm to obtain the results of the second visualization scheme in this article. The third visualization scheme is used to directly map the attention of the classification token to the feature tokens back to the input image. Our standard model has 12 transformer encoder blocks, and each MSA has four heads. The attention of each MSA is combined by the following equations:

$$A^{(i)} = \sum_{j=1}^{4} A^{ij}, i = 1, 2, \cdots, 12 \tag{28}$$

$$A = \sum_{i=1}^{12} \frac{A^{(i)} - \min\left(A^{(i)}\right)}{\max\left(A^{(i)}\right) - \min\left(A^{(i)}\right)} \tag{29}$$

where A^{ij} denotes the attention map of classification token to feature tokens in j-th head of i-th transformer encoder, and $A^{(i)}$ is the attention map that fuses the attention maps of all the heads in i-th transformer encoder. A will be mapped directly to the input image. This visualization scheme does not involve any gradient calculation. It will reflect the attention of the classification token to feature tokens and demonstrate whether the calculation of MSA without increasing parameters is effective for identifying diseased maize leaves.

Evaluation of Model Performance

We chose accuracy, precision, recall, F1 score, parameters, and floating-point operations per second (FLOPs) to evaluate our classification model. Among them, precision, recall, and F1 score can be calculated by the following equations:

$$Precision = \frac{TP}{TP + FP} \tag{30}$$

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{31}$$

$$F1 = \frac{2Precision \times \mathrm{Re}call}{Precision + \mathrm{Re}call} \tag{32}$$

where TP refers to the number of true positives, FP refers to the number of false positives, and FN refers to the number of false negatives.

RESULTS

All models in this article were trained with 110 epochs. **Figure 6** shows the accuracy and loss of all models as a function of epochs. As can be seen, the performance dramatically improves within the first 20 epochs, but improvement is minor beyond 20 epochs. We compared five mainstream CNNs with our standard model. These CNNs have achieved excellent performance on some specific tasks. For example, MobileNet (Sandler et al., 2018) can be applied to mobile terminals due to its lightweight architecture. ResNet (He et al., 2016) as a baseline is widely used in the industry. EfficientNet (Tan and Le, 2019) has a relatively significant advantage in terms of speed and accuracy. **Table 2** compares the standard model with these CNNs in terms of six metrics (i.e., accuracy, precision, recall, F1 score, parameters, and FLOPs). The accuracies of CNNs reached VGG11 (Simonyan and Zisserman, 2014) 97.9%, ResNet50 96.6%, EfficientNet-b3 91.6%, Inception-v3 (Szegedy et al., 2016) 97.2%, and MobileNet-v2-140 90.7%, whereas the standard model reached 98.7% and surpassed these CNNs. **Figure 7** shows that comparison of the accuracy trends of the standard model with the mainstream CNNs and ViT-base during training. The recall of the standard model for class H is 1% lower than that of Vgg11, but it surpasses Vgg11 in all other metrics. Except for VGG11, the standard model surpasses or ties the rest of these CNNs in accuracy, precision, recall, and F1 score. For the FLOPs metric, MobileNet-v2-140 has lower FLOPs than the standard model and requires less computing power. Since MobileNet-v2-140 is designed for mobile terminals, its FLOPs must be lower than common models. Nevertheless, MobileNet-v2-140 has 6.6 times the number of the standard model parameters. The number of parameters and FLOPs of other models are significantly higher

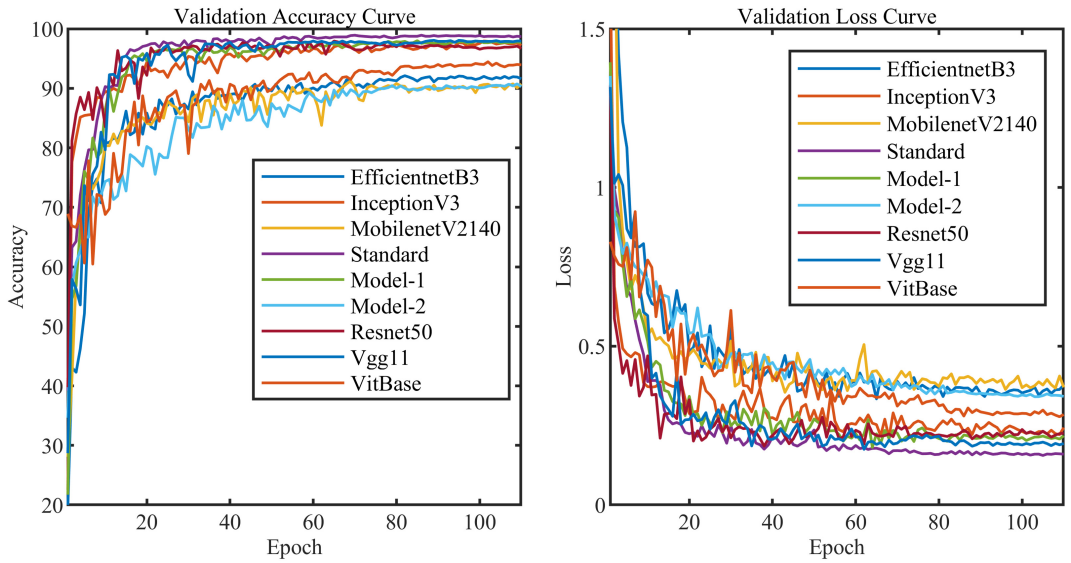

FIGURE 6 | Accuracy curve and loss curve of validation set of all models in this article (i.e., Model-1 and Model-2).

TABLE 2 | Comparison between the standard model and the other mainstream models.

	Standard	VGG11	EfficientNet-b3	Inception-v3	MobileNet-v2-140	ResNet50	Vit-base
Accuracy (%)	98.7	97.9	91.6	97.2	90.2	96.6	93.9
Precision (%)							
H	97	96	88	96	88	94	91
SCLB	99	99	90	97	88	99	92
SR	99	98	94	98	92	96	98
GLS	100	100	97	99	99	100	96
Recall (%)							
H	99	100	92	98	93	99	95
SCLB	97	96	86	94	86	91	90
SR	100	99	98	100	96	99	97
GLS	99	97	89	98	85	97	92
F1 (%)							
H	98	98	90	97	91	97	93
SCLB	98	97	88	95	87	95	91
SR	100	98	96	99	94	98	98
GLS	100	98	93	99	91	98	94
Parameter (M)	0.65	128.78	10.70	21.79	4.32	23.52	82.80
FLOPs (G)	1.47	7.61	1.62	5.72	0.59	4.10	17.58

than the standard model (**Figure 8** clearly shows the comparison of parameters and FLOPs of the models), which means that these CNNs are designed to be bloated for maize leaf disease identification in a complex background. As can be seen, for the specific task of this article, stacking the number of layers of the network and increasing the number of parameters of the model are not effective in improving the performance of the model. Our model has only one convolutional layer and one pooling layer to encode local regions of images into tokens, and transformer encoder as the core computational module, which not only significantly reduces the number of parameters and FLOPs of the model but also achieves the best performance. From another perspective, although the number of parameters of the standard model is on average three orders of magnitude lower than the other models in **Table 2**, its FLOPs are in the same order of magnitude as theirs. Since MSA involves large-scale matrix computation when computing the attention matrix between tokens, this operation does not involve the model's parameters but increases the model computation. A comparison in **Table 2** between the standard and ViT (accuracy 93.9%) was created to compare the patch embedding method with the convolution method, showing that the convolution method is superior to the patch embedding method from the perspective of results, which indicates that convolutional layer and max-pooling

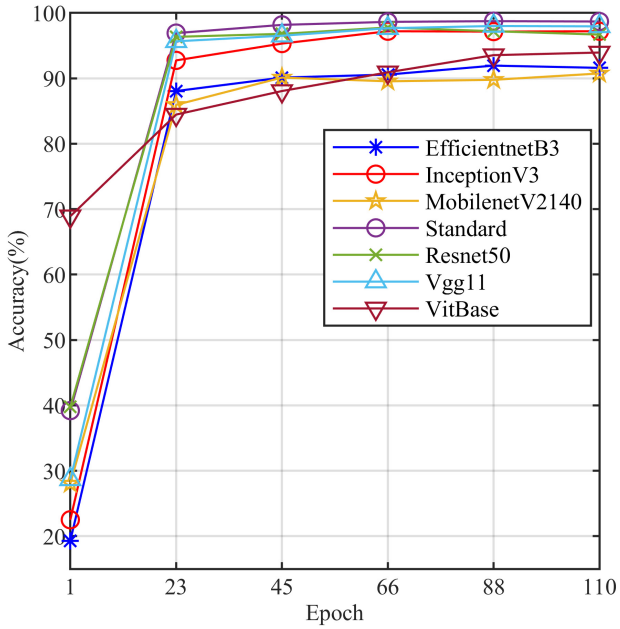

FIGURE 7 | The accuracy line chart of the standard model and the other models.

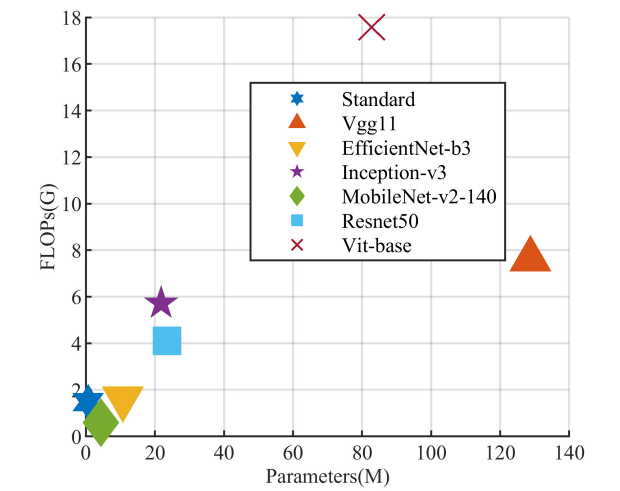

FIGURE 8 | Comparison of the standard model and the other models in parameters and FLOPs.

layer can sufficiently encode information of maize leaf disease lesions into tokens and reduce model's parameters.

Figure 9 shows the confusion matrices of all models of this article (i.e., Model-1 and Model-2). The confusion matrix's abscissa axis represents actual class and ordinate axis represents predicted class. As can be seen, for the nine models, they always tend to identify the SCLB class as the H class. SCLB lesions on maize leaves are minor and scattered, which results in some samples infected similar to H class. In contrast, considering computing power limitation, the size of images can be shrunk small, which leads to SLCB lesions-pixels disappearing and classification error. SR and GLS are rarely misclassified, because their symptoms are markedly distinct from other categories of this article. SR lesions on the leaf tissue's aboveground surface resemble flecks that develop into small golden-brown pustules or bumps. Tan lesions of SR can be distinguished readily from yellow lesions on the surface of maize leaf infected SCLB or GLS.

Table 3 compares the three models to explore the necessity of the self-attention. Model-1 and Model-2 (**Figure 5**) are modified from the standard model to conduct this study. Model-1 fuses feature tokens into a classification token in Stage 3 by a linear layer instead of adding a classification token at the end of Stage 1, and Model-2 removes the MSA based on Model-1. **Figure 10** clearly shows the increased curve of accuracy of the three models. The accuracy of the standard model exceeds Model-1 by 1%. They have almost the same number of parameters, which indicates that the classification token participating in MSA computation is better than fusing feature tokens into classification tokens. The accuracy of Model-2 is substantially lower than Model-1 by 7.5%. Among other metrics (e.g., precision, recall, and F1 score), Model-2 is also substantially lower than Model-1. The expected results indicate that the self-attention dramatically improves the performance of the model. Model-1 and Model-2 have the same number of parameters, but the FLOPs of Model-2 are much lower than those of Model-1. As mentioned above, the large-scale matrix operations involved in MSA do not increase the number of parameters in the model but do increase the computational complexity of the model. This little computational cost is worth the significant improvement it brings to the model, which also shows that self-attention, a computation that involves almost no parameters of the model, can dramatically improve the identification of maize leaf diseases in complex backgrounds.

In addition, we compared the effect of different train and validation set ratios on the accuracy of the standard model (**Table 4**). As can be seen, the model's accuracy gradually increases as the ratio increases. When the ratio reaches 20–80%, the accuracy reaches 94.0%, while when the ratio reaches 50–50%, the accuracy almost stops increasing. **Figure 11** shows the validation accuracy curve of the standard model over 9 ratios in the training process. The experiment indicates that the standard model can achieve satisfactory performance even when the number of training samples is small.

Figure 12 provides the results of the visualization of the regions of interest to the model during the classification process. We chose ResNet50 to compare with the standard model and three visualization schemes. For the convolutional or pooling layer-based scheme, we chose the output of the last convolutional layer of layer2 of ResNet50 and the output of the first pooling layer of the standard model because they both output feature maps with a width of 28. In the tokens-based visualization scheme, we selected the output of the first LN layer in the last transformer encoder of the standard model. In the attention matrix-based visualization scheme, we combined the attention matrix of all transformer encoders in the entire model. By comparing **Figures 12A,B**, as can be seen, in field settings with complex backgrounds, ResNet50 has a large amount of attention scattered in the background. In contrast, the attention

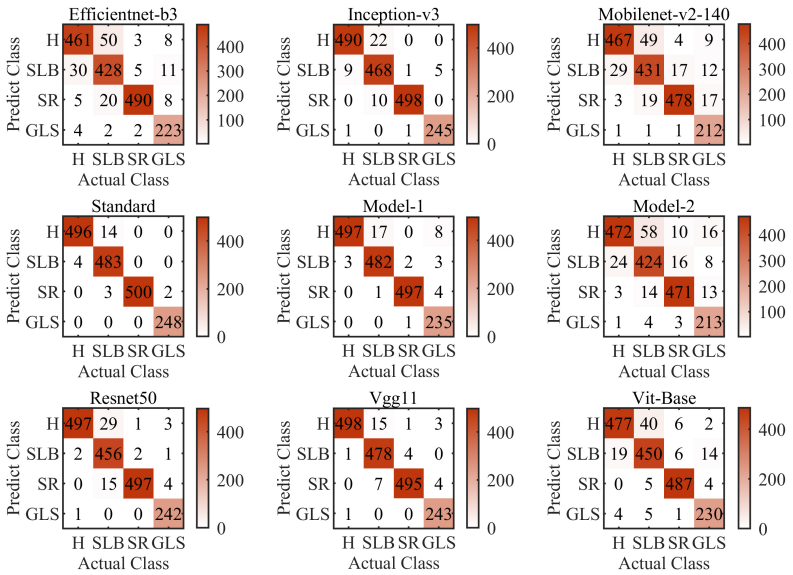

FIGURE 9 | Confusion matrices of all models of this article (i.e., Model-1 and Model-2).

TABLE 3 | Research of importance of the self-attention.

	Standard	Model-1	Model-2
Accuracy (%)	98.7	97.7	90.2
Precision (%)			
H	97	95	85
SCLB	99	98	90
SR	99	99	94
GLS	100	100	96
Recall (%)			
H	99	99	94
SCLB	97	96	85
SR	100	99	94
GLS	99	94	85
F1 (%)			
H	98	97	89
SCLB	98	97	87
SR	100	99	94
GLS	100	97	90
Parameter (M)	0.65	0.66	0.66
FLOPs (G)	1.47	1.46	0.33

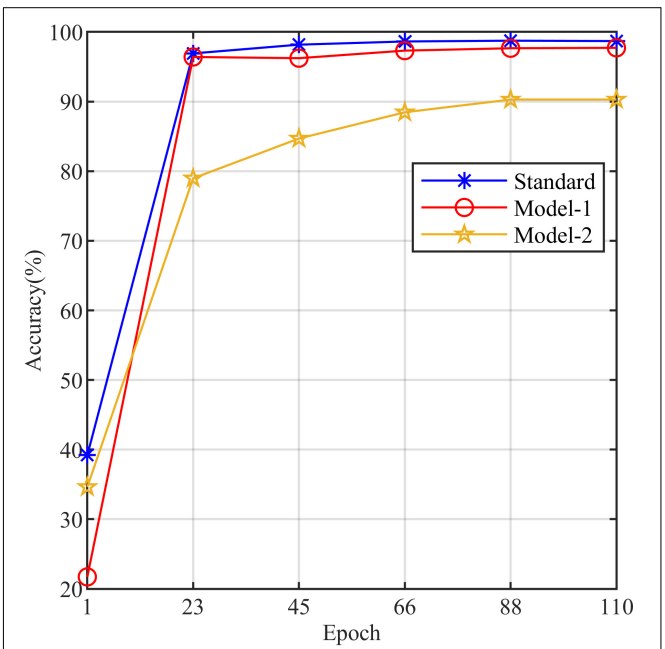

FIGURE 10 | The accuracy line chart of the standard model, Model-1, and Model-2.

of the standard model is mainly focused on the leaf surface. **Figure 12C** shows that the attention is more refined when representing features based on tokens, effectively suppressing the background information and focusing more on the leaf surface lesions. **Figure 12D** shows the attention distribution of classification token to other feature tokens, which is consistent with the area of attention of the model, which also shows that the MSA calculation mechanism that does not increase the number of model parameters effectively enhances the attention of the model to crucial information and suppresses the useless background noise information.

DISCUSSION

The common CNNs represent the feature information of an image *via* feature maps, and deepening the depth of the network can generally achieve better performance, but this also increases the number of model parameters and computational effort. They are more suitable for object recognition. The pixels where these objects are located usually do not have similarities, and the overall

TABLE 4 | The standard model accuracy results for each train-validation set.

Train-test split (%)	H	SCLB	SR	GLS	Accuracy
10–90	243/2,192	224/2,019	202/1,821	100/900	0.894
20–80	487/1,948	448/1,795	404/1,619	200/800	0.940
30–70	730/1,705	672/1,571	606/1,417	300/700	0.967
40–60	974/1,461	897/1,346	809/1,214	400/600	0.980
50–50	1,217/1,218	1,121/1,122	1,011/1,012	500/500	0.977
60–40	1,461/974	1,345/898	1,213/810	600/400	0.980
70–30	1,704/731	1,570/673	1,416/607	700/300	0.986
80–20	1,948/487	1,794/449	1,618/405	800/200	0.990
90–10	2,191/244	2,018/225	1,820/203	900/100	0.989
Total	2,435	2,243	2,023	1,000	

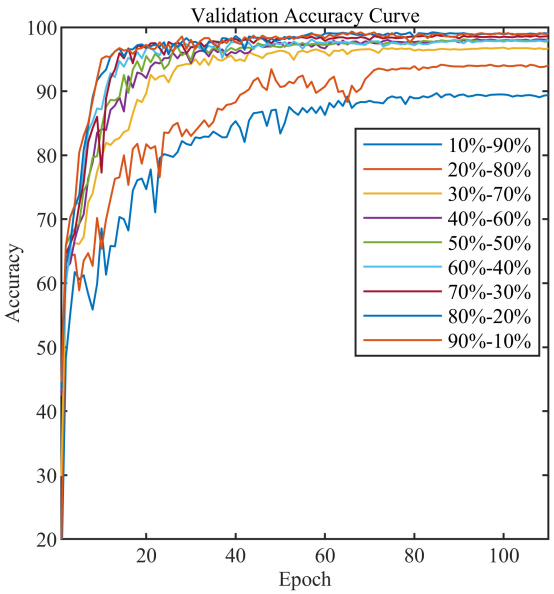

FIGURE 11 | The validation accuracy curve of the standard model in nine train-validation sets.

pixels composition of the pattern presents the features of the target object. For maize leaf disease recognition, the pixels where the lesions are located usually have similarities (reflected in the RGB values), which requires a feature representation with higher resolution rather than the feature maps of CNNs. Since the feature maps increases with the number of channels but decreases in width as the network feeds forward. The relationship between lesions information and receptive field becomes blurred. There is no correlation computed between the lesions, so increasing the number of network layers will only bring a slight increase in recognition rate while also increasing the volume and complexity of the network. The model used in this article is entirely different from CNNs in that it is based on tokens to represent the visual information of local areas of the image. Stage 1 encodes the visual information of the receptive field into a matrix of feature tokens. The subsequent network does not perform any compression of this matrix. However, it continuously computes the correlation (attention) between tokens by MSA, making the network pay more attention to information about the lesions useful for classification and suppressing the noisy information in the background. We demonstrated this idea from this article's theoretical, experimental, and visual analysis perspectives. Tokens represent the local feature information of images,

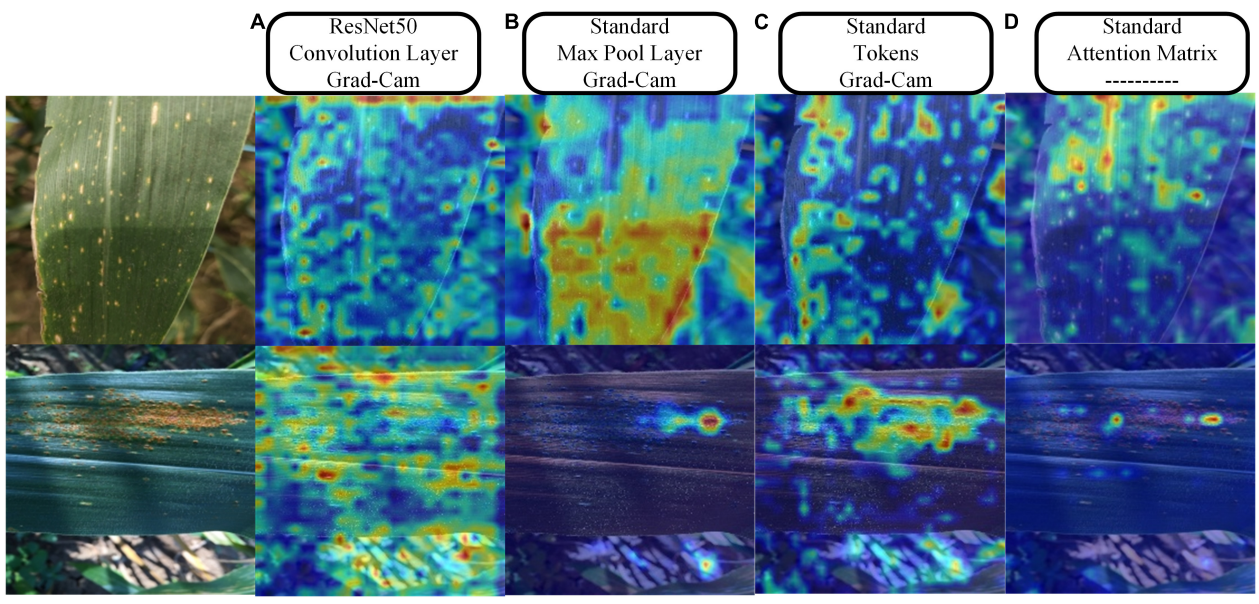

FIGURE 12 | Visualization results of the three schemes. **(A)** Grad-Cam method for visualization of the feature maps output by the Resnet50 convolutional layer. **(B)** Grad-Cam method for visualization of the feature maps output from the max-pooling layer of the standard model. **(C)** Grad-Cam method for visualization of the tokens of the standard model. **(D)** Directly map the attention of the classification token on feature tokens to the original image.

and self-attention calculates the correlation of local information, which is more suitable for maize leaf disease identification in complex background. Therefore, guided by the above analysis, we designed a more reasonable model that achieves the best performance with minimal computational cost and number of parameters compared with other mainstream CNNs. However, our model has some limitations. The token (i.e., a single vector) dimension is a hyperparameter. As it increases, the feature information can be represented more abundantly, increasing the attention matrix's scale. Large-scale matrix operations can rapidly increase the computational complexity of the model. Many researchers are now actively working to overcome this challenge (Carion et al., 2020; Liu et al., 2021; Touvron et al., 2021).

In addition, the results above indicate that convolution method outperforms the patch embedding method in encoding maize disease features into feature tokens. Convolution kernel as receptive field extracts visual information by sliding of itself. Two slides of the receptive field have an overlapped area, associating the semantic information of the area. However, the patch embedding method cuts a complete image into many irrelevant patches and directly encodes these patches into tokens, leading to the semantic information of adjacent areas to be lost. Humans tend to process critical vision information instead of all receptive field information, which is mainly limited by the brain's inability to process massive information simultaneously. The mechanism by which humans process visual information is consistent with our model based on the attention mechanism, and they both prefer critical information.

In the field of plant disease identification, the hyperspectral imaging technology is usually used for object detection because the difference in reflectance of plant disease features is slight (Yue et al., 2015; Polder et al., 2019; Wang D. et al., 2019). The investigation of Nagasubramanian et al. (2019) demonstrated that soybeans infected the charcoal rot are more sensitive than healthy soybeans in the wavelengths of visible spectra (400–700 nm). Yang et al. (2021) have achieved good results in the Citrus Huanglongbing detection task by fusing hyperspectral data in CNNs using a multimodal approach. Recent research has shown that the transformer architecture is better suited for multimodal tasks (Frank et al., 2021; Zhang et al., 2021). We will conduct research by extending our model to combine with multimodal approaches for crop disease identification and detection in complex backgrounds in future.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

XQ and KL conceived the study and wrote the manuscript. XQ implemented the algorithm. LC and CZ described the diseases and provided the dataset. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Open-End Fund of Information Materials and Intelligent Sensing Laboratory of Anhui Province (IMIS202009) and Anhui Agricultural University Introduction and Stabilization of Talents Research Funding Project (No. yj2020-74).

REFERENCES

Ahila Priyadharshini, R., Arivazhagan, S., Arun, M., and Mirnalini, A. (2019). Maize leaf disease classification using deep convolutional neural networks. *Neural Comput. Applic.* 31, 8887–8895. doi: 10.1007/s00521-019-04228-3

Aregbesola, E., Ortega-Beltran, A., Falade, T., Jonathan, G., Hearne, S., and Bandyopadhyay, R. (2020). A detached leaf assay to rapidly screen for resistance of maize to Bipolaris maydis, the causal agent of southern corn leaf blight. *Eur. J. Plant Pathol.* 156, 133–145. doi: 10.1007/s10658-019-01870-4

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1607.06450 (accessed April 1, 2021).

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). "End-to-end object detection with transformers," in *Proceedings of the European Conference on Computer Vision*, (Cham: Springer), 213–229. doi: 10.1007/978-3-030-58452-8_13

Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. (2018). "Grad-cam++: generalized gradient-based visual explanations for deep convolutional networks," in *Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, (Piscataway, NJ: IEEE), 839–847. doi: 10.1109/WACV.2018.00097

Chen, L., Chen, J., Hajimirsadeghi, H., and Mori, G. (2020). "Adapting Grad-CAM for embedding networks," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, (Piscataway, NJ: IEEE), 2794–2803. doi: 10.1109/WACV45572.2020.9093461

DeChant, C., Wiesner-Hanks, T., Chen, S., Stewart, E. L., Yosinski, J., Gore, M. A., et al. (2017). Automated identification of northern leaf blight-infected maize plants from field imagery using deep learning. *Phytopathology* 107, 1426–1432. doi: 10.1094/PHYTO-11-16-0417-R

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1810.04805 (accessed March 25, 2021).

Dhaka, V. S., Meena, S. V., Rani, G., Sinwar, D., Kavita, Ijaz, M. F., et al. (2021). A survey of deep convolutional neural networks applied for prediction of plant leaf diseases. *Sensors* 21:4749. doi: 10.3390/s21144749

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: transformers for image recognition at scale. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/2010.11929 (accessed May 22, 2021).

Frank, S., Bugliarello, E., and Elliott, D. (2021). Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/2109.04448 (accessed February 22, 2022).

Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., and Wang, Y. (2021). Transformer in transformer. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/2103.00112 (accessed July 31, 2021).

Hassani, A., Walton, S., Shah, N., Abuduweili, A., Li, J., and Shi, H. (2021). Escaping the big data paradigm with compact transformers. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/2104.05704 (accessed August 21, 2021).

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Piscataway, NJ: IEEE), 770–778. doi: 10.1109/CVPR.2016.90

Hendrycks, D., and Gimpel, K. (2016). Gaussian error linear units (gelus). *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1606.08415 (accessed August 21, 2021).

Jiang, P.-T., Zhang, C.-B., Hou, Q., Cheng, M.-M., and Wei, Y. (2021). Layercam: exploring hierarchical class activation maps for localization. *IEEE Trans. Image Process.* 30, 5875–5888. doi: 10.1109/TIP.2021.3089943

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Adv. Neural Inform. Process. Syst.* 25, 1097–1105.

Kundu, N., Rani, G., Dhaka, V. S., Gupta, K., Nayak, S. C., Verma, S., et al. (2021). IoT and interpretable machine learning based framework for disease prediction in pearl millet. *Sensors* 21:5386. doi: 10.3390/s21165386

Li, Y., Yang, L., Xu, B., Wang, J., and Lin, H. (2019). Improving user attribute classification with text and social network attention. *Cogn. Comput.* 11, 459–468. doi: 10.1007/s12559-019-9624-y

Liu, J., and Wang, X. (2021). Plant diseases and pests detection based on deep learning: a review. *Plant Methods* 17, 1–18. doi: 10.1186/s13007-021-00722-9

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). Swin transformer: hierarchical vision transformer using shifted windows. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/2103.14030 (accessed September 26, 2021). doi: 10.1109/ICCV48922.2021.00986

Lu, J., Xiong, C., Parikh, D., and Socher, R. (2017). "Knowing when to look: adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Piscataway, NJ: IEEE), 375–383. doi: 10.1109/CVPR.2017.345

Lu, Y., Yi, S., Zeng, N., Liu, Y., and Zhang, Y. (2017). Identification of rice diseases using deep convolutional neural networks. *Neurocomputing* 267, 378–384. doi: 10.1016/j.neucom.2017.06.023

Lv, M., Zhou, G., He, M., Chen, A., Zhang, W., and Hu, Y. (2020). Maize leaf disease identification based on feature enhancement and DMS-robust alexnet. *IEEE Access* 8, 57952–57966. doi: 10.1109/access.2020.2982443

Nagasubramanian, K., Jones, S., Singh, A. K., Sarkar, S., Singh, A., and Ganapathysubramanian, B. (2019). Plant disease identification using explainable 3D deep learning on hyperspectral images. *Plant Methods* 15, 1–10. doi: 10.1186/s13007-019-0479-8

Ouppaphan, P. (2017). "Corn disease identification from leaf images using convolutional neural networks," in *Proceedings of the 2017 21st International Computer Science and Engineering Conference (ICSEC)*, (Piscataway, NJ: IEEE), 1–5. doi: 10.1109/ICSEC.2017.8443919

Panigrahi, K. P., Das, H., Sahoo, A. K., and Moharana, S. C. (2020). "Maize leaf disease detection and classification using machine learning algorithms," in *Progress in Computing, Analytics and Networking*, eds P. K. Pattnaik, S. S. Rautaray, H. Das, and J. Nayak (Cham: Springer), 659–669. doi: 10.1155/2022/6504616

Polder, G., Blok, P. M., de Villiers, H. A., van der Wolf, J. M., and Kamp, J. (2019). Potato virus Y detection in seed potatoes using deep learning on hyperspectral images. *Front. Plant Sci.* 10:209. doi: 10.3389/fpls.2019.00209

Ranum, P., Peña-Rosas, J. P., and Garcia-Casal, M. N. (2014). Global maize production, utilization, and consumption. *Ann. N.Y. Acad. Sci.* 1312, 105–112. doi: 10.1111/nyas.12396

Saito, B. C., Silva, L. Q., Andrade, J. A. C., and Goodman, M. M. (2018). Adaptability and stability of corn inbred lines regarding resistance to gray leaf spot and northern leaf blight. *Crop Breed. Appl. Biotechnol.* 18, 148–154. doi: 10.1590/1984-70332018v18n2a21

Saleem, M. H., Potgieter, J., and Mahmood Arif, K. (2019). Plant disease detection and classification by deep learning. *Plants* 8:468. doi: 10.3390/plants8110468

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). "Mobilenetv2: inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Piscataway, NJ: IEEE), 4510–4520. doi: 10.1109/CVPR.2018.00474

Savary, S., Ficke, A., Aubertot, J.-N., and Hollier, C. (2012). Crop losses due to diseases and their implications for global food production losses and food security. *Food Security* 4, 519–537. doi: 10.1007/s12571-012-0200-5

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). "Grad-cam: visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, (Piscataway, NJ: IEEE), 618–626. doi: 10.1109/ICCV.2017.74

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1409.1556 (accessed September 1, 2021). doi: 10.3390/s21082852

Song, S., Lan, C., Xing, J., Zeng, W., and Liu, J. (2017). "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, (Piscataway, NJ: IEEE).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Piscataway, NJ: IEEE), 2818–2826. doi: 10.1109/CVPR.2016.308

Tan, M., and Le, Q. (2019). "Efficientnet: rethinking model scaling for convolutional neural networks," in *Proceedings of the International Conference on Machine Learning: PMLR*, (Piscataway, NJ: IEEE), 6105–6114.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). "Training data-efficient image transformers & distillation through attention," in *Proceedings of the International Conference on Machine Learning: PMLR*, (Piscataway, NJ: IEEE), 10347–10357.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems*, eds M. S. Kearns, S. A. Solla, and D. A. Cohn (Cambridge, MA: MIT Press), 5998–6008.

Wang, D., Vinson, R., Holmes, M., Seibel, G., Bechar, A., Nof, S., et al. (2019). Early detection of tomato spotted wilt virus by hyperspectral imaging and outlier removal auxiliary classifier generative adversarial nets (OR-AC-GAN). *Sci. Rep.* 9, 1–14. doi: 10.1038/s41598-019-40066-y

Wang, S., Chen, Z., Tian, L., Ding, Y., Zhang, J., Zhou, J., et al. (2019). Comparative proteomics combined with analyses of transgenic plants reveal Zm REM 1.3 mediates maize resistance to southern corn rust. *Plant Biotechnol. J.* 17, 2153–2168. doi: 10.1111/pbi.13129

Wieczorek, M., Sika, J., Wozniak, M., Garg, S., and Hassan, M. (2021). "Lightweight CNN model for human face detection in risk situations," in *Proceedings of the IEEE Transactions on Industrial Informatics*, (Piscataway, NJ: IEEE). doi: 10.1109/TII.2021.3129629

Yang, D., Wang, F., Hu, Y., Lan, Y., and Deng, X. (2021). Citrus huanglongbing detection based on multi-modal feature fusion learning. *Front. Plant Sci.* 12:809506. doi: 10.3389/fpls.2021.809506

Yang, G., He, Y., Yang, Y., and Xu, B. (2020). Fine-grained image classification for crop disease based on attention mechanism. *Front. Plant Sci.* 11:600854. doi: 10.3389/fpls.2020.600854

Yue, J., Zhao, W., Mao, S., and Liu, H. (2015). Spectral–spatial classification of hyperspectral images using deep convolutional neural networks. *Remote Sens. Lett.* 6, 468–477. doi: 10.1080/2150704x.2015.1047045

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. (2019). "Cutmix: regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (Piscataway, NJ: IEEE), 6023–6032. doi: 10.1109/ICCV.2019.00612

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2018). "mixup: beyond empirical risk minimization," in *Proceedings of the International Conference on Learning Representations*, (Piscataway, NJ: IEEE).

Zhang, L. N., and Yang, B. (2014). Research on recognition of maize disease based on mobile internet and support vector machine technique. *Adv. Mater. Res.* 905, 659–662. doi: 10.4028/www.scientific.net/amr.905.659

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., et al. (2021). "Vinvl: revisiting visual representations in vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (Piscataway, NJ: IEEE), 5579–5588. doi: 10.1109/CVPR46437.2021.00553

Zhang, Z., He, X., Sun, X., Guo, L., Wang, J., and Wang, F. (2015). Image recognition of maize leaf disease based on GA-SVM. *Chem. Eng. Trans.* 46, 199–204.

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.