Image Processing, Electronics and Computers
L. Trajkovic et al. (Eds.)

doi:10.3233/ATDE240471

Swin Transformer Based on Image Enhancement Algorithm

Liwei CHEN[a], Gulinazi AILIMUJIANG[b,1] and Zhichuang ZHAO[a]
[a] *School of Electronic and Engineering, Yili Normal University, Yining, Xinjiang, China*
[b] *School of Network Security and Information Technology, Yili Normal University, Yining, Xinjiang, China*

Abstract. In the field of plant taxonomy, species often exhibit high levels of similarity in morphological features, color expression, and surface textures, while also containing rich and complex detail information. These characteristics pose significant challenges in the identification and classification process. Traditional machine learning cannot extract features comprehensively and accurately. This study leverages the Swin Transformer combined with image enhancement algorithms for plant image classification. On one hand, it benefits from enlarging the inter-class distance to improve classification accuracy; on the other hand, it addresses the issue of high computational complexity in large-scale plant image processing. By integrating the Swin Transformer with advanced image enhancement techniques in the task of plant image classification, a significant performance improvement has been achieved. Compared to using the Swin Transformer method alone, this integrated strategy has shown superior results, achieving an accuracy of 89.03% in plant classification tasks. This paper focuses on the plant image classification process based on the Swin Transformer with image enhancement algorithms.

Keywords. Swin Transformer; Shifted Windows; Plant Image Classification; Edge Detection; Data Augmentation

1. Introduction

Plant image classification is a significant research direction in the field of computer vision, which can help relevant fields better understand and utilize the diverse plant resources. Many scholars have made outstanding contributions to this field, yet there are still some problems in practical applications. Therefore, this paper aims to propose a new method that achieves high classification accuracy and practicality, providing a more effective solution for the application of plant image classification. In the current field of plant classification, models such as Convolutional Neural Networks[1], AlexNet[2], VGG[3-5], GoogLeNet[6,7], and ResNet[8-10]are commonly used. Agarwal[11] et al. proposed a CNN model for plant detection using leaf images, which classifies 10 types of leaf images using 3 convolutional layers, 3 max pooling layers, and 2 fully connected layers. Experimental data show that compared with pretrained models like VGG16 (77.20%) and MobileNet[12](63.75%), Agarwal's method achieved the highest test

[1] Corresponding Author: Gulinazi AILIMUJIANG, 1014818024@qq.com.

accuracy of 91.20%. Guo et al.[13] proposed a multi-scale detection convolutional neural network model, which can better extract plant features, thereby improving the model's accuracy. However, existing research indicates that the limited receptive field of convolution at each layer makes it difficult for CNNs to capture long-distance dependency information. To address this issue, Vision Transformer[14-17] (ViT) uses the Multi-headed Self-attention (MHSA) mechanism to increase the diversity of context information in multiple projection subspaces, thus enhancing feature representation capabilities. The Swin Transformer, based on the ViT model concept, indicates that the quality of input images directly affects model performance in plant image classification tasks. Therefore, certain preprocessing of input data is necessary in practical applications to enhance model performance. This study first uses an edge detection algorithm to obtain the edge information of input images and merges the color edge images with the original images to highlight the contours of large objects. Subsequently, adaptive histogram equalization techniques are used to enhance image colors, improving image quality through increased contrast, brightness, and saturation. This can effectively improve the information quality and discriminative power of images, achieving an accuracy of 89.03% in plant classification tasks.

2. Relevant Theory

2.1 Swin Transformer

The ViT has shown outstanding performance in many competitions and tasks within the field of computer vision. The structure of the ViT is shown in Figure 1.

Building on the ideas behind the ViT model, the Swin Transformer[18-22] innovatively introduces a sliding window mechanism. This allows the model to learn information across windows while downsampling enables the processing of super-resolution images, saving computational resources and allowing the model to focus on both global and local information. The structure of the Swin Transformer is shown in Figure 2.

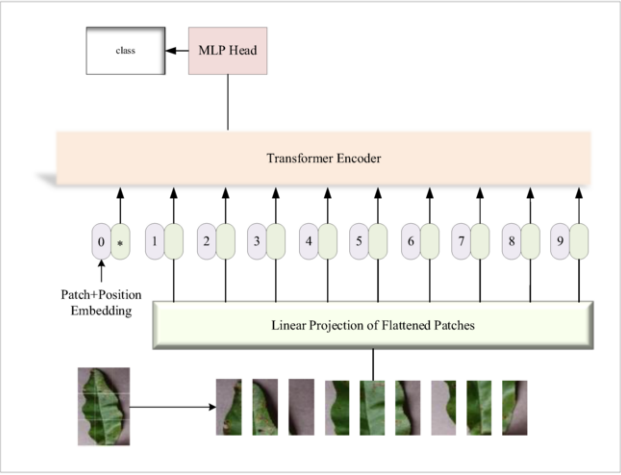

Figure 1. Vision Transformer architecture diagram

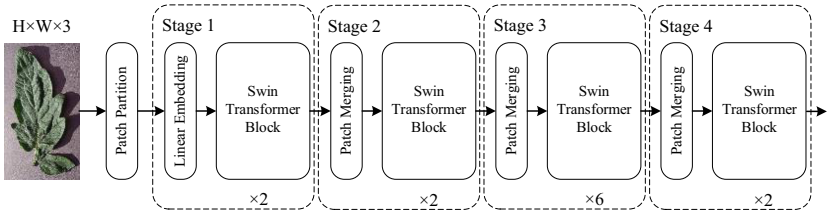

Figure 2. Swin Transformer architecture diagram

The Swin Transformer adopts a hierarchical structure similar to that of convolutional neural networks, dividing the feature map into several non-overlapping windows, and computing Multi-Head Self-Attention within each window. Compared to the ViT model, the Swin Transformer significantly reduces computational demands.Swin Transformer Blocks introduce the Window Multi-Head Self-Attention (W-MSA) module and the Shifted Window Multi-Head Self-Attention (SW-MSA) module to reduce computation. The Multi-Head Self-Attention (MSA) module in traditional architectures requires calculating self-attention for each pixel with every other pixel across the feature map, leading to substantial computational overhead. In contrast, the Swin Transformer partitions the feature map into windows and computes self-attention within these windows, dramatically decreasing the amount of computation. The difference in computational demand is illustrated in formulas (1) and (2).

$$\Omega(MSA) = 4hwC^2 + 2(hw)^2C \tag{1}$$

$$\Omega(W-MSA) = 4hwC^2 + 2M^2hw \tag{2}$$

Among them, h represents the height of the feature map, w represents the width of the feature map, C represents the depth of the feature map, and M represents the size of each window.

In the Transformer Block structure, first perform Windows Multi-head Self-Attention (W-MSA), followed by Shifted Windows Multi-Head Self-Attention (SW-MSA). The computations of consecutive Swin Transformer Blocks are shown in equations (3), (4), (5), and (6).

$$\hat{Z}^l = W-MSA(LN(Z^{l-1})) + Z^{l-1} \tag{3}$$

$$Z^l = MLP(LN(\hat{Z}^l)) + \hat{Z}^l \tag{4}$$

$$\hat{Z}^{l+1} = SW-MSA(LN(Z^l)) + Z^l \tag{5}$$

$$Z^{l+1} = MLP(LN(\hat{Z}^{l+1})) + \hat{Z}^{l+1} \tag{6}$$

Here, $\hat{Z}^l$andZ^l respectively represent the output features of blockl for the (S) W-MSA module and the MLP module.

2.2 Image Enhancement Algorithm

The study employs Gaussian blur to smooth the image by calculating the average of surrounding pixels at each pixel point to reduce noise. Gaussian blur is shown in Equation (7).

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x-\mu)^2+(y-\nu)^2}{2\sigma^2}} \tag{7}$$

G(x, y) represents the weight at the center position of the filter, where x and y are the offsets of pixels within the filter relative to the center. μ and ν denote the central position of the Gaussian function, and σ is the standard deviation controlling the width of the Gaussian function (i.e., the degree of blurriness).

The direction of the gradient is perpendicular to the direction of the edge. The edge detection operator returns Gx in the horizontal direction and Gy in the vertical direction. The magnitude G and direction θ of the gradient are shown in equations (8) and (9) as follows:

$$G = \sqrt{G_x + G_y} \tag{8}$$

$$\theta = arctan\frac{G_y}{G_x} \tag{9}$$

After obtaining the magnitude and direction of the gradient, traverse the pixels in the image to remove all non-edge points.

This study uses an image fusion formula to blend the original image with the edge image, as shown in equation (10):

$$F(i,j) = \alpha \times A(i,j) + \beta \times B(i,j) \tag{10}$$

Here, α and β are the fusion weight parameters used to control edge strength. Subsequently, the Contrast Limited Adaptive Histogram Equalization (CLAHE)[23] algorithm is utilized to enhance the contrast of the luminance channel. The CLAHE algorithm is a type of adaptive histogram equalization technique that enhances contrast by dividing the image into small blocks and performing histogram equalization on each block. Finally, the image with enhanced contrast is color adjusted using the HSV[24] color space. The HSV color space consists of Hue, Saturation, and Value components. By adjusting the values of hue, brightness, and saturation, the color and vividness of the image are altered. Lastly, a contour detection algorithm is applied to the edge image for highlighting the contours of large objects. By calculating the area of the contours, those with an area greater than a threshold are selected, and contour lines are drawn on the original image.

2.3 Swin Transformer Based On Image Enhancement Algorithm

When using Swin Transformer for plant image classification tasks, the quality of the input images will directly affect the performance of the model. Therefore, in practical applications, it is necessary to preprocess the input data to some extent in order to improve model performance.

This study applies techniques such as edge detection, color enhancement, and highlighting the edges of large objects. By combining them, it can effectively improve the image's information quality and discriminative ability. The study first uses an edge detection algorithm to obtain the edge information of the input image and merges the color edge image with the original image to highlight the contours of large objects. Then, it uses adaptive histogram equalization technology to enhance the image color by improving image quality through increased contrast, brightness, and saturation. To verify the effectiveness of this method, experimental tests were conducted on the CIFAR-10 and plant datasets. The Swin Transformer model was used as the basis for the experiments to compare the impact of different preprocessing methods on model performance. When using the Swin Transformer model for image classification tasks, the image preprocessing method proposed in this study can effectively improve the model's performance and generalization ability. Future research directions include optimizing algorithms, adding new data augmentation algorithms, etc., to further explore the advantages of the Swin Transformer model and improve its performance and reliability in practical applications. The Swin Transformer structure based on image enhancement is shown in Figure 3.

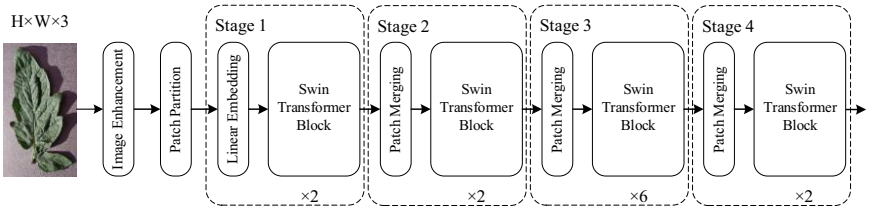

Figure 3. Structure Diagram Of Swin Transformer Based On Image Enhancement Algorithms

3. Experimental Analysis

3.1 Experimental Dataset

The dataset contains 3,410 samples, divided into 31 different plant categories, including fruits, vegetables, flowers, etc. Each category has 60 images. The shooting angle and background of each sample are different, and some samples are partially obscured by shadows. Part of the dataset's images is shown in Figure 4.

Figure 4. Partial plant images from the dataset

The dataset comprises 1,488 images in the training set, accounting for 44%, 372 images in the validation set, accounting for 11%, and 1,550 images in the test set, accounting for 45%. Each sample exhibits different characteristics of plants. This study uses the plant dataset to accomplish classification tasks. By training and testing the dataset, the performance of the methods proposed in this study is evaluated. At the same time, comparisons with publicly available datasets are made to provide supportive evidence for the results of this study. The dataset selects economically significant crops with a wide variety, contributing to future analyses of crops. Before using the dataset, this study conducted some data preprocessing. First, the experimental dataset was cleaned to remove some anomalous images. Then, feature selection was performed on the experimental dataset, choosing features relevant to this study's experiments.

3.2 Parameter Settings

This study conducts pre-training on the ImageNet-1K[25] dataset, using the obtained weights as the initial weights for the plant classification task. On this basis, the plant dataset, consisting of 1,488 initial training samples, is trained using the Swin Transformer model. In this paper, the model is trained for 100 epochs with a batch size of 4, utilizing the AdamW optimizer with a learning rate of 0.0001. To mitigate overfitting, weight decay is set to 0.05.

3.3 Evaluating Indicator

In the comparative experiments, this study primarily uses Accuracy as the evaluation metric to compare and determine the effectiveness of the improved methods proposed in this research.

Accuracy is the most commonly used performance metric for classification. It represents the precision of the model, which is the number of correctly identified instances divided by the total number of instances. Generally, the higher the accuracy of a model, the better its performance is considered to be. The formula for accuracy is as shown in Equation (11).

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \tag{11}$$

3.4 Analysis of Experimental Results

The experiment compared the accuracy of the Swin Transformer and the Swin Transformer based on image augmentation algorithms as shown in Tables 1 and 2. It was found that accuracy improved, indicating that the improved Swin Transformer is more suitable for plant classification and exhibits higher robustness. This study utilized the CIFAR-10 classification task and a plant classification task encompassing 31 categories as experimental benchmarks. In the experiments, the Swin Transformer based on image augmentation algorithms and the Swin Transformer were applied to the CIFAR-10 classification task and the plant classification task, respectively, using the same parameter settings. The model architecture, optimizer, and hyperparameters used in the experiments were consistent. The experimental results for the CIFAR-10 dataset and the plant classification task are shown in Tables 1 and 2.

Initially, this study compares the classification accuracy between the Swin Transformer and the Swin Transformer based on image augmentation algorithms in plant classification tasks. After averaging the results of multiple experiments, the accuracy of the Swin Transformer was found to be 88.32%, while the Swin Transformer based on image augmentation algorithms achieved an accuracy of 89.03%. Compared to the original Swin Transformer, the Swin Transformer based on image augmentation algorithms achieved a performance improvement of 0.71% in accuracy. This indicates that the Swin Transformer enhanced by image augmentation algorithms has achieved significant effectiveness in classification tasks.

Table 1. Experimental results

Method	Accuracy
Swin Transformer	94.40%
Swin Transformer Based On Image Enhancement Algorithm	95.01%

Table 2. Experimental results

Method	Accuracy
Swin Transformer	88.32%
Swin Transformer Based On Image Enhancement Algorithm	89.03%

To more comprehensively evaluate the performance differences between the Swin Transformer and the Swin Transformer based on image augmentation algorithms, this study presents their multi-class confusion matrices. By observing the confusion matrices, one can understand the classification performance of the Swin Transformer based on image augmentation algorithms across different categories. The study noted that, in some specific categories, the Swin Transformer based on image augmentation algorithms achieved better classification results compared to the original Swin Transformer. There were also some classification errors in certain categories, providing a focus and direction for further optimizing the Swin Transformer based on image augmentation algorithms. The confusion matrices are shown in Figures 5 and 6.

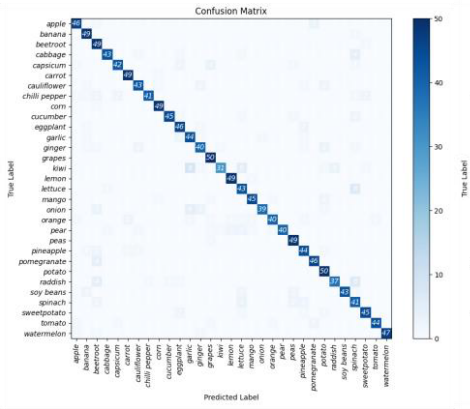

Figure 5. Confusion matrix for Swin Transformer

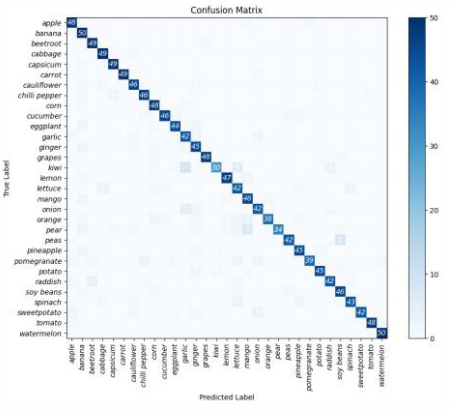

Figure 6. Improved confusion matrix

This study evaluates the training and generalization performance of the model by plotting the training loss curve, validation loss curve, training accuracy curve, and validation accuracy curve. Ideally, both training and validation losses should decrease over time, while training and validation accuracies should increase, without too significant a gap between them. Observing Figures 7 and 9 reveals that both the training accuracy curve and the validation accuracy curve gradually rise with the increase in iteration numbers.The calculated variances for the Swin Transformer model are as follows: the training loss curve variance is 0.0164, the training accuracy curve variance is 0.0013, the validation loss curve variance is 0.0086, and the validation accuracy curve variance is 0.0004. For the Swin Transformer model based on image augmentation algorithms, the training loss curve variance is 0.0186, the training accuracy curve variance is 0.0014, the validation loss curve variance is 0.0068, and the validation accuracy curve variance is 0.0003. The comparison suggests that the Swin Transformer model based on image enhancement has smaller variances in both validation loss and accuracy curves, indicating better generalization performance. The loss and accuracy curves are depicted in Figures 7, 8, 9, and 10.

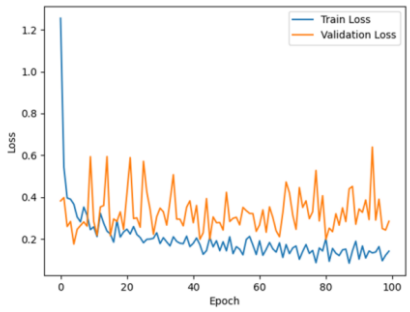

Figure 7. Loss curve for Swin Transformer

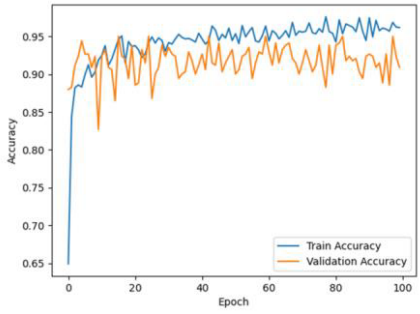

Figure 8. Accuracy curve for Swin Transformer

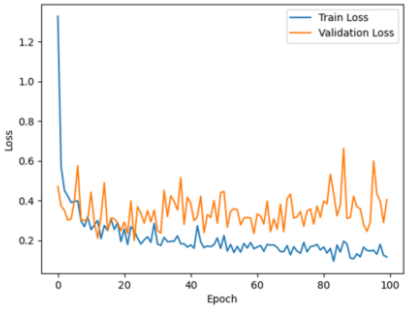

Figure 9. Improved loss curve

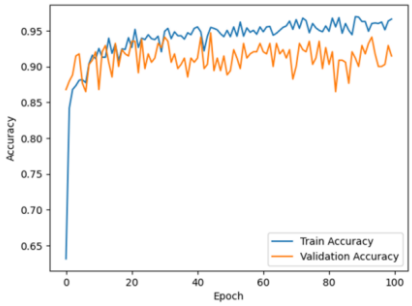

Figure 10. Improved accuracy curve

The experiment compared the classification process of plant images, finding that the Swin Transformer model enhanced with image augmentation algorithms achieved higher accuracy in complex images. For example, as shown in the first image of Figure 11, a chili pepper partially obscured by leaves and differing in color, where green peppers blend in with the background making them hard to distinguish, and the presence of flowers and leaves in the image could influence the classification outcome. The Swin Transformer model classified this image as a chili pepper with a 94% probability, whereas the model enhanced with image augmentation algorithms classified it as a chili pepper with a 99.9% probability.In the second image, the corn includes various growth stages, leading to a diversity of shapes and textures in the corn plants. Due to the dense and closely packed corn leaves, occlusion between leaves often occurs, preventing some of the corn plants from being fully visible. Additionally, variations in lighting in the image, particularly strong frontal lighting altering brightness and contrast, pose challenges for classification. The Swin Transformer model classified this image as corn with an 80.2% probability, while the model enhanced with image augmentation algorithms did so with a 100% probability, indicating it was not influenced by the external environment.The third image was taken from a low angle, causing the shape and features of the bananas to appear distorted. This angle might lead to a different appearance from traditional views. Due to the low-angle shot, the upper part of the image is brighter, and shadows and reflections have changed. The background includes the sky, other plants, and dead leaves, which may have colors and textures similar to the bananas, making it difficult for the algorithm to distinguish the bananas from the background. The fact that the bananas are seen from below and are connected makes their shape appear different, adding to the classification challenge. The Swin Transformer model misclassified this image, while the model based on image enhancement algorithms classified it as bananas with an 88.4% probability.In the fourth image, the lemons are round and slightly flattened, with varying sizes. The lemon's skin has a delicate texture, presenting an uneven appearance, which means texture features can be easily influenced by lighting conditions. The presence of occluding objects makes part of the lemon not fully visible, adding to the classification difficulty. The Swin Transformer model classified this image as a lemon with a 96.6% probability, whereas the model enhanced with image augmentation algorithms did so with a 99.2% probability. Through the experiments, it was determined that the accuracy of classification by the Swin Transformer model based on image enhancement algorithms was relatively improved. The classification process for some of the images is shown in Figure 11.

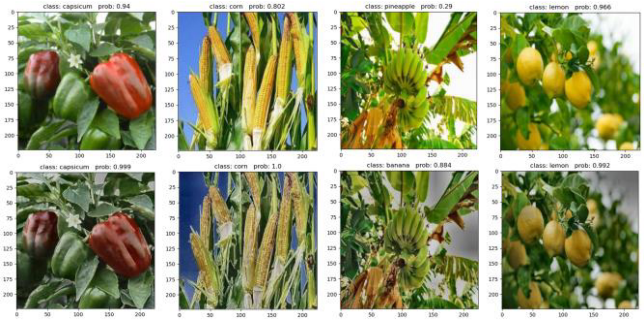

Figure 11. Classification results of partial images in the dataset

4. Conclusion

Plant classification based on deep learning methods has been widely applied and achieved good engineering application results. However, the quality of the input images directly affects the model's performance, leading to the proposal of the Swin Transformer algorithm based on image enhancement. In comparative analyses, this study observed that the Swin Transformer achieved an accuracy of 88.32% in plant classification tasks, while the Swin Transformer based on image enhancement achieved an accuracy of 89.03%, indicating a performance improvement of 0.71%. This demonstrates the effectiveness of the improved method in plant classification tasks. Nonetheless, this study also recognizes some shortcomings, possibly due to the characteristics of the dataset or limitations of the model. Therefore, the next steps involve further improving the model and optimizing algorithms to increase classification accuracy.

References

[1] ZHANG K, ZUO W, GU S, et al. Learning deep CNN denoiser prior for image restoration//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 3929-3938.

[2] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 2012, 25.

[3] SIMONYAN K, ZISSERMAN A. Verydeep convolutional networks for large-scaleimage recognition. arXiv:1409.1556, 2014.

[4] ISMAIL FAWAZ H, LUCAS B, FORESTIER G, et al. Inceptiontime: Finding alexnet for time series classification. Data Mining and Knowledge Discovery, 2020, 34(6): 1936-1962.

[5] BALLESTER P, ARAUJO R. On the performance of GoogLeNet and AlexNet applied to sketches//Proceedings of the AAAI conference on artificial intelligence. 2016, 30(1).

[6] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[7] YUESHENG F, JIAN S, FUXIANG X, et al. Circular fruit and vegetable classification based on optimized GoogLeNet. IEEE Access, 2021, 9: 113599-113611.

[8] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[9] TARG S, ALMEIDA D, LYMAN K. Resnet in resnet: Generalizing residual architectures.arXiv:1603.08029, 2016.

[10] WU Z, SHEN C, VAN DEN HENGEL A. Wider or deeper: Revisiting the resnet model for visual recognition. Pattern Recognition, 2019, 90: 119-133.

[11] A M A, B A S, C S A, et al. ToLeD: Tomato Leaf Disease Detection using Convolution Neural Network. Procedia Computer Science, 2020, 167:293-301.

[12] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.

[13] Guo X Q, Fan T J, Shu X. Tomato leaf diseases recognition based on improved Multi-Scale AlexNet. Transactions of the Chinese Society of Agricultural Engineering, 2019, 35(13): 162-169 (inChinese)

[14] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929, 2020.

[15] WANG W, XIE E, LI X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions//Proceedings of the IEEE/CVF international conference on computer vision.2021:568-578.

[16] KHAN S, NASEER M, HAYAT M, et al. Transformers in vision: A survey. ACM computing surveys (CSUR), 2022, 54(10s): 1-41.

[17] CHEN C F R, FAN Q, PANDA R. Crossvit: Cross-attention multi-scale vision transformer for image classification//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 357-366.

[18] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows//Proceedings of the IEEE/CVF international conference on computer vision.2021:10012-10022.

[19] LIANG J, CAO J, SUN G, et al. Swinir: Image restoration using swin transformer//Proceedings of the IEEE/CVF international conference on computer vision, 2021: 1833-1844.

[20] HATAMIZADEH A, NATH V, TANGY, et al. Swin unetr: Swin transformers forsemantic segmentation of brain tumors in mri images//International MICCAI Brainlesion Workshop. Cham: Springer International Publishing, 2021: 272-284.

[21] HE X, ZHOU Y, ZHAO J, et al. Swin transformer embedding UNet for remote sensing image semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60:1-15.

[22] LIU Z, TAN Y, HE Q, et al. SwinNet: Swin transformer drives edge-aware RGB-D and RGB-T salient object detection. IEEE Transactions on Circuits and Systemsfor Video Technology, 2021, 32(7):4486-4497.

[23] REZA A M. Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement. Journal of VLSI signal processing systems for signal, image and video technology, 2004, 38: 35-44.

[24] SURAL S, QIAN G, PRAMANIK S. Segmentation and histogram generation using the HSV color space for image retrieval//Proceedings. International Conference on Image Processing. IEEE, 2002, 2: II-II

[25] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge. International journal of computer vision, 2015, 115: 211-252.