agriculture

MDPI

Article

Potato Plant Variety Identification Study Based on Improved Swin Transformer

Xue Xing [1], Chengzhong Liu [1,*], Junying Han [1], Quan Feng [2], Enfang Qi [3], Yaying Qu [3] and Baixiong Ma [1]

[1] College of Information Sciences and Technology, Gansu Agricultural University, Lanzhou 730070, China; xingxgs@sina.com (X.X.); hanjy@gsau.edu.cn (J.H.); mabaixiong1998@163.com (B.M.)
[2] College of Mechanical and Electrical Engineering, Gansu Agricultural University, Lanzhou 730070, China; fengq@gsau.edu.cn
[3] Potato Research Institute, Gansu Academy of Agricultural Sciences, Lanzhou 730070, China; qefang@126.com (E.Q.); moxyu000@163.com (Y.Q.)
* Correspondence: liucz@gsau.edu.cn; Tel.: +86-138-9321-8815

Abstract: Potato is one of the most important food crops in the world and occupies a crucial position in China's agricultural development. Due to the large number of potato varieties and the phenomenon of variety mixing, the development of the potato industry is seriously affected. Therefore, accurate identification of potato varieties is a key link to promote the development of the potato industry. Deep learning technology is used to identify potato varieties with good accuracy, but there are relatively few related studies. Thus, this paper introduces an enhanced Swin Transformer classification model named MSR-SwinT (Multi-scale residual Swin Transformer). The model employs a multi-scale feature fusion module in place of patch partitioning and linear embedding. This approach effectively extracts features of various scales and enhances the model's feature extraction capability. Additionally, the residual learning strategy is integrated into the Swin Transformer block, effectively addressing the issue of gradient disappearance and enabling the model to capture complex features more effectively. The model can better capture complex features. The enhanced MSR-SwinT model is validated using the potato plant dataset, demonstrating strong performance in potato plant image recognition with an accuracy of 94.64%. This represents an improvement of 3.02 percentage points compared to the original Swin Transformer model. Experimental evidence shows that the improved model performs better and generalizes better, providing a more effective solution for potato variety identification.

Keywords: deep learning; potato; Swin Transformer; convolutional neural network; variety identification

check for updates

Academic Editors: Gniewko Niedbała and Francesco Marinello

Received: 19 October 2024
Revised: 11 December 2024
Accepted: 27 December 2024
Published: 2 January 2025

Citation: Xing, X.; Liu, C.; Han, J.; Feng, Q.; Qi, E.; Qu, Y.; Ma, B. Potato Plant Variety Identification Study Based on Improved Swin Transformer. *Agriculture* **2025**, *15*, 87. https://doi.org/10.3390/agriculture15010087

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Potato (*Solanum tuberosum* L.), as a tuber crop, has become one of the world's most important food staples due to its strong adaptability and high nutritional value. It is ranked among the seven major global food crops, following rice, maize, and wheat [1–3] Advances in agricultural technology have led to the continuous emergence of improved potato varieties, enhancing yields and disease resistance, thus opening new horizons for agricultural production. However, this progress also introduces challenges in accurately identifying new varieties to ensure their safety and high quality, crucial for maintaining food security.

Traditionally, potato variety identification has relied on methods such as morphological feature observation, physio-biochemical analysis, and molecular marker technique [4–6].

These approaches are time-consuming, susceptible to subjective biases, and involve complex procedures. In recent years, the application of machine learning algorithms [7] in variety identification has enhanced recognition efficiency. However, these methods are limited by the constraints of manually designed feature extraction and the high demand for computational resources [8–10].

Advancements in artificial intelligence, particularly the application of convolutional neural networks (CNNs) in image classification, have enabled automatic feature extraction, significantly enhancing classification accuracy and efficiency [11–13]. For instance, Liu et al. [14] utilized the TensorFlow framework to construct a deep CNN for classifying 14 apple varieties, achieving an overall accuracy of 97.11%. Gao et al. [15] developed the CMPNet model, integrating ResNet-50, SE-ReNet, and SE-ReNeXt through transfer learning and Concat layers, which improved wheat variety identification accuracy to 99.51% with an inference time of just 0.0212 s. Laabassi et al. [16] employed transfer learning to train five CNN architectures for identifying four wheat grain varieties, with DenseNet201, Inception V3, and MobileNet achieving the highest test accuracies of 95.68%, 95.62%, and 95.49%, respectively. Guo [17] improved an optimized lightweight classification model based on the ShuffleNetV2 network, and this model demonstrated excellent performance in potato plant recognition, effectively enhancing the recognition accuracy and efficiency.

In June 2017, Google introduced the Transformer model in their paper "Attention is All You Need" [18], which rapidly garnered significant attention in the academic community. Starting in 2018, researchers began applying Transformer models to image recognition, leading to extensive research in this area [19–21]. Reedha R et al. [22] proposed a weed image classification method combining Vision Transformer (ViT) and transfer learning, focusing on drone-captured images of beet, parsley, and spinach fields. Their experiments demonstrated that ViT outperformed state-of-the-art models like EfficientNet and ResNet, even with small-scale annotated training datasets. Borhani Y et al. [23] developed a lightweight deep-learning approach based on ViT for automatic plant disease classification. By comparing classical convolutional neural networks (CNNs) with hybrid CNN-ViT models, they found that while attention blocks improved accuracy, they reduced prediction speed. Combining attention blocks with CNN layers mitigated this speed issue. Luo [24] proposed an improved Swin Transformer model for potato plant recognition, which achieved good recognition results.

The aforementioned studies highlight that, despite the excellent performance of convolutional neural networks (CNNs) and Vision Transformers in plant image classification, potato variety identification remains challenging. Complex field backgrounds increase the difficulty of image recognition, while tubers growing underground necessitate inferring feature information indirectly from above-ground plants. Additionally, there is a lack of sufficient research and technological accumulation in this area. To address these challenges, this paper proposes the MSR-SwinT model, which integrates the local feature extraction advantages of CNNs with the long-range dependency modeling capabilities of Swin Transformers. This fusion significantly enhances feature representation and classification performance for potato plants, providing robust technical support for intelligent agricultural management. The primary contributions of this paper are outlined below:

(1) A total of 69 varieties of potato plant images were collected at the Yuzhong Experimental Station. After rigorous screening and organization, five datasets with varying sample sizes were constructed for diverse experimental and analytical purposes.
(2) We propose and optimize a novel model, MSR-SwinT, based on the Swin Transformer architecture. This model enhances the capture of complex features and incorporates residual learning strategies to improve training stability and learning efficiency.

(3) Through comparative training and validation against classical network models on our self-built potato plant dataset, experimental results demonstrate the superior performance and effectiveness of the MSR-SwinT model in classification tasks.

2. Materials and Methods

2.1. Potato Plants Data Acquisition and Data Preprocessing

The potato leaf data for this study were obtained from the Potato Experimental Station at Yuzhong Horticultural Farm, Gansu Provincial Academy of Agricultural Sciences. The station's coordinates are 35.85° N latitude and 104.12° E longitude, with an elevation of 1960 m above sea level. Its climatic characteristics consist of a mean annual temperature of 6.7 °C, average annual precipitation of 400 mm, and a frost-free period lasting around 120 days, indicating a temperate semi-arid climate. Data collection employed a Nikon COOLPIX B700 digital camera, set to automatic mode with an ISO maximum of 1600 and a minimum shutter speed of 1/30 s, ensuring the capture of clear images. All photographs were taken under natural lighting conditions, and the potato plants were photographed from different angles and saved in JPG format at a resolution of 2272 × 1704.

The images were gathered during the period spanning from 25 July 2023 to 1 August 2023, which included sunny weather, 2 days of light rain, and 1–2 days of cloudy weather; this weather variability helped to capture leaf images in different light conditions. To increase the diversity of the data, different periods, such as morning, midday, and evening, were chosen for the photographs. Due to the complexity and irregularity of the growth pattern of potato plants, such as growth attitude, branching status, and growth direction, four images with a single background and two images with a natural background were taken for each potato plant in the dataset to make the model more effective at recognizing potato plants in their natural state. Figure 1 shows some examples of potato plants.



Figure 1. Selected samples of potato plants, where (**a**–**d**) are blue background images and (**e**,**f**) are natural background images.

This study collected five potato plant datasets of varying scales, named data_150, data_100, data_70, data_50, and data_40, collectively covering 69 different potato varieties. Specifically, data_150 includes 30 varieties, with approximately 150 plants per variety, totaling 27,300 images; data_100 includes 10 varieties, with approximately 100 plants per variety, totaling 6336 images; data_70 includes 9 varieties, with approximately 70 plants per variety, totaling 3912 images; data_50 includes 10 varieties, with approximately 50 plants per

variety, totaling 3174 images; data_40 includes 10 varieties, with approximately 40 plants per variety, totaling 2484 images. The number of varieties (lines) and each species in each dataset is shown in Table 1.

Table 1. Varieties and numbers of each data set.

Name of Data Set	Variety	Number	Variety	Number	Variety	Number
data_150	L11146-2	876	Ly1503-9	912	Ly1628-1	912
	LS-19	888	L11120-8	912	L13104-12	912
	LS-20	876	LS-12	930	Ly1615-3	924
	LS-23	876	Ly1624-1	918	Ly1619-3	930
	LS-25	864	Ly1602-3	918	Ly1733-16	924
	Ly1503-10	882	Ly1521-2	918	L11133-2H	906
	LS-17	852	Ly1520-9	918	Ly1702-3	930
	LS-26	864	Ly1722-1	924	Ly1819-8	936
	Ly1624-2	888	Ly1727-3	1020	Ly1544-2	960
	L14140-5	912	L12116-19	906	Ly1517-8	912
data_100	Ly1742-2	648	Ly1826-8	618	Ly1725-1	642
	Ly1728-12	636	Ly1929-1	648	Ly1945-1	636
	Ly1826-4	624	Ly1620-1	624		
	Ly1815-9	618	Ly1962-1	642		
data_70	Ly1912-1	444	Ly1907-20	438	Ly1946-2	444
	Ly1864-1	438	Ly1185-1	432	Ly1956-1	426
	Ly1820-2	432	Ly1843-1	426	Ly1525-3	432
data_50	Ly1818-3	342	Ly1819-2	312	Ly1832-2	324
	Ly1907-26	312	Ly1841-2	306	Ly1879-5	318
	Ly1908-19	312	Ly19108-7	330		
	Ly1844-2	306	Ly1841-1	312		
data_40	Ly19105-6	246	Ly1860-3	252	Ly1826-2	246
	Ly1963-2	252	Ly1860-9	252	Ly1627-50	246
	Ly19123-1	264	Ly19120-3	252	Ly19123-1	264
	Ly1740-6	246	Ly1631-1	228	Ly1740-6	246

When confronted with the challenge of limited training data, the model is often prone to overfitting, i.e., it has learned the training data in too much detail, which will lead to its weakened generalization ability when dealing with new data. Data expansion introduces more diverse training samples to effectively mitigate the overfitting problem. This enables the model to learn from a broader array of samples, thereby enhancing its ability to generalize when encountering unfamiliar data. In addition, by simulating various changes in the real-world environment, such as differences in lighting and angles, data expansion can also help the model adapt to more complex and changing real-world scenarios. For the datasets with relatively small amounts of data, namely data_100, data_70, data_50, and data_40, a series of data augmentation techniques were implemented. Specifically, in terms of brightness and contrast adjustment, an operation with a brightness adjustment factor of 1.1 was applied to the images with a probability of 50% to simulate situations under different illumination intensities. In addition, a 180° clockwise rotation operation was performed on all images. Through the comprehensive application of these methods, the scale of each dataset was successfully expanded to three times its initial size, providing a richer data foundation for subsequent model training and contributing to the improvement of the model's performance.

The individual datasets were partitioned into training, validation, and test sets at an 8:1:1 ratio. To improve model training efficiency and recognition performance, all images were scaled according to the network's input pixel requirements. In the training set, the

input images were first randomly cropped and then scaled to the standard size of 224 × 224 pixels required by the network. Next, these images were randomly flipped horizontally and randomly flipped vertically. Finally, all images that underwent flipping were normalized using the transforms. Normalize ([0.5, 0.5, 0.5], [0.5, 0.5, 0.5]) function. Specifically, for each pixel value in each channel of the image, the transformation is performed according to Equation (1):

$$x_{new} = \frac{x - mean}{std} \quad (1)$$

where x is the original pixel value and x_{new} is the normalized pixel value. The mean and standard deviation for each channel are set to 0.5. This normalization process standardizes the pixel values, making the data distribution more conducive to model training and thereby enhancing the model's performance in potato variety identification tasks. The validation set and the test set use the same data processing method, where random cropping and scaling to 224 × 224 pixels are performed, followed by normalization. An example of data preprocessing is shown in Figure 2.

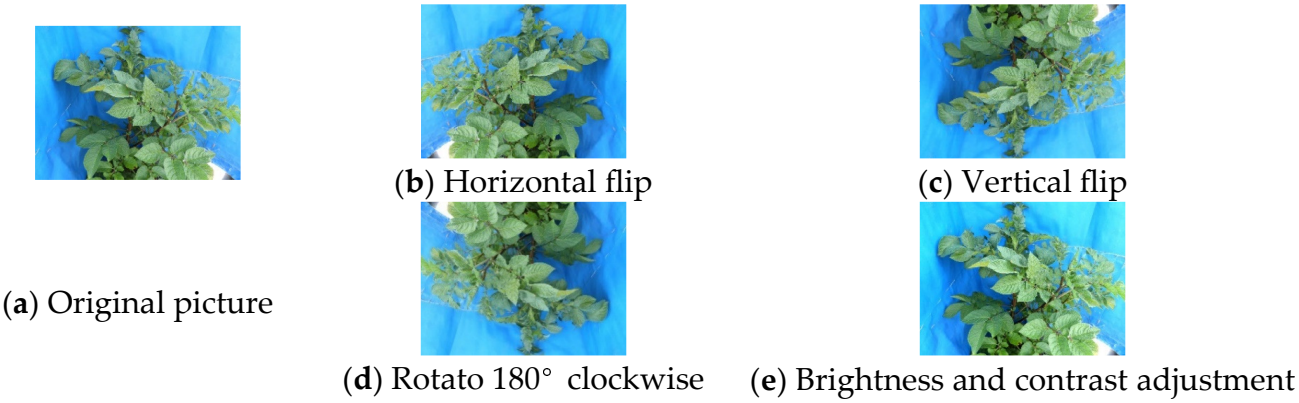


Figure 2. Example of data preprocessing.

2.2. Swin Transformer

Swin Transformer [25] is an improved model based on the Transformer architecture, specifically designed for computer vision tasks. It excels in handling image data, particularly in applications such as object detection and image segmentation.

Swin Transformer uses a hierarchical structure similar to the pyramid structure found in traditional convolutional neural networks. The overall structure is shown in Figure 3, and its detailed configuration is shown in Table 2. The network is divided into tiers that progressively reduce the feature map and increase the channel count.

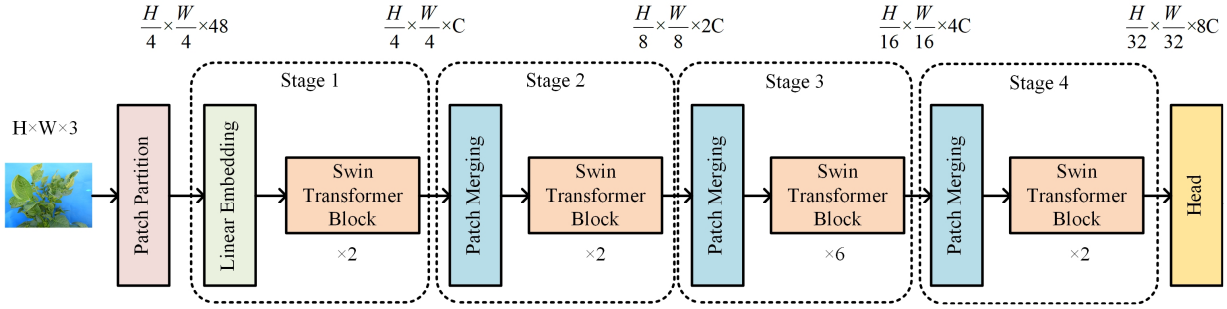


Figure 3. Swin Transformer model structure.

Table 2. Detailed configuration of Swin Transformer.

Output Size	Module Name	Configuration
56 × 56	Patch Partition Linear Embedding	Concat 4 × 4, 96-d, LayerNorm
	Swin Transformer Block	$\begin{bmatrix}\text{window size } 7\times 7\\ \text{dim} = 96, \text{head} = 3\end{bmatrix} \times 2$
28 × 28	Patch Merging	Concat 2 × 2, 192-d, LayerNorm
	Swin Transformer Block	$\begin{bmatrix}\text{window size } 7\times 7\\ \text{dim} = 192, \text{head} = 6\end{bmatrix} \times 2$
14 × 14	Patch Merging	Concat 2 × 2, 384-d, LayerNorm
	Swin Transformer Block	$\begin{bmatrix}\text{window size } 7\times 7\\ \text{dim} = 384, \text{head} = 12\end{bmatrix} \times 2$
7 × 7	Patch Merging	Concat 2 × 2, 768-d, LayerNorm
	Swin Transformer Block	$\begin{bmatrix}\text{window size } 7\times 7\\ \text{dim} = 768, \text{head} = 24\end{bmatrix} \times 2$
1 × 1	Head	AdaptiveAvgPool, Flatten, Linear

The input image is first partitioned into patches using Patch Partition, where each 4 × 4 block of adjacent pixels is treated as a single patch and flattened along the channel dimension. These patches then undergo a linear transformation via a Linear Embedding layer; the detailed process is shown in Figure 4. Subsequently, the model constructs feature maps of varying resolutions through four stages. Each stage includes a Patch Merging layer for downsampling, which reduces the spatial dimensions (width and height) by half while doubling the number of channels. Figure 5 illustrates the patch merging process.

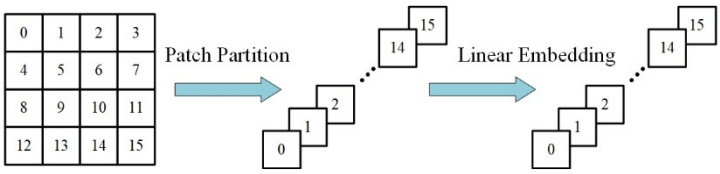


Figure 4. Patch Partition and Linear Embedding process.

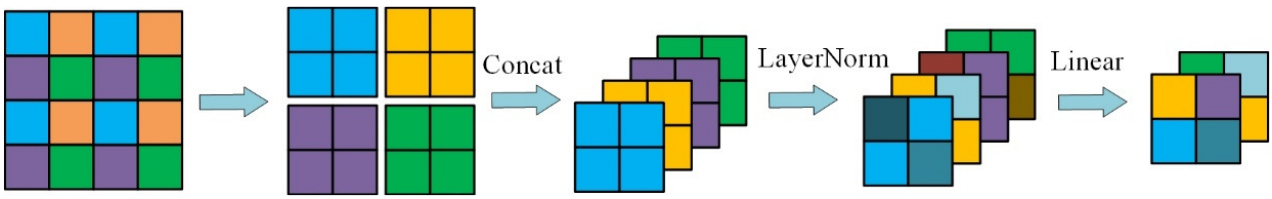


Figure 5. Patch Merging operation process.

Following the Patch Merging layer, each stage stacks multiple Swin Transformer Blocks. These blocks consist of two complementary structures: Window-based Multi-head Self-Attention (W-MSA) and Shifted Window Multi-head Self-Attention (SW-MSA). The W-MSA mechanism divides the feature map into smaller windows to reduce computational complexity, while the SW-MSA mechanism achieves information exchange by offsetting window positions, thereby enhancing the model's ability to capture global information. The architecture of the Swin Transformer is illustrated in Figure 6.

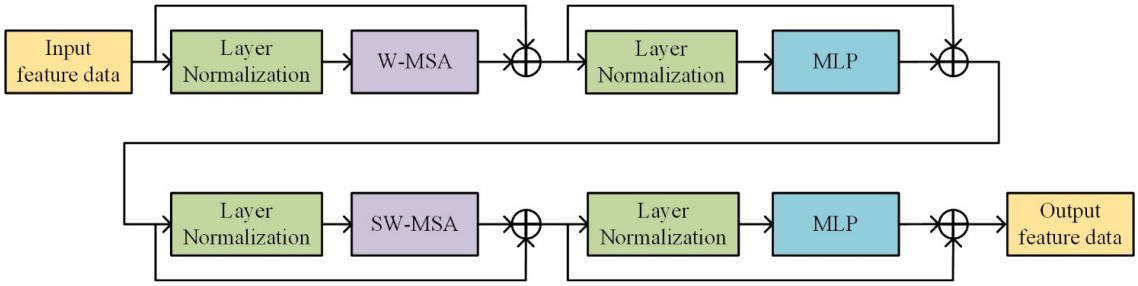


Figure 6. Swin Transformer Block structure.

Each element must interact with all other elements in the feature map during the Self-Attention calculation in the standard Multi-Head Self-Attention (MSA) module.

The W-MSA module employs a distinct method: it initially partitions the feature map into several small, non-overlapping windows and then conducts the Self-Attention computation separately within each window. This approach significantly reduces computational complexity because each pixel only interacts with pixels within the window it is in, thus reducing the amount of Self-Attention that needs to be computed. The computational complexity of each MSA and W-MSA are as shown in (Equations (2) and (3)):

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

$$\Omega(W-MSA) = 4hwC^2 + 2M^2hwC \tag{3}$$

In (Equations (1) and (2)), h is the height of the feature map, w is the width of the feature map, C is the depth of the feature map, and M is the width and height of each window.

With the W-MSA module, self-attention calculation occurs within each window independently, restricting information exchange between different windows. To overcome this limitation, Swin Transformer introduces the SW-MSA module, which enables information transfer between adjacent windows by dislocating (shifting) the windows so that information can be shared between them, thereby enhancing the model's capacity to capture global information.

2.3. Multi-Scale Feature Fusion Module

The multi-scale feature fusion module is designed with inspiration from the Inception module [26], which emphasizes the integration of features across multiple levels and scales. The Inception module excels at processing features of various scales simultaneously within a network. Inspired by this idea, the paper introduces a multi-scale feature fusion module. This module is crafted to capture and blend image features across diverse levels and scales, enhancing the model's capability to recognize and represent potato plant features within intricate scenes. Different scales of features in an image can be captured when the image is convolved using different sizes of convolution kernels. For example, smaller convolution kernels (e.g., 1×1) help to achieve cross-channel information fusion, while larger convolution kernels (e.g., 5×5, 7×7) are able to cover a larger receptive field and capture more complex image structures. By sensibly combining these convolutional filters of varying sizes, different local features in the image can be effectively extracted, resulting in a richer and more comprehensive feature representation.

Different potato plants have different characteristics, including different shape profiles, growth potentials, and branching structures. Traditional fixed scale convolutional kernels have limitations in capturing these complex features. To capture these detailed differences, a Multi-Scale Feature Fusion (MSFF) module was utilized in this study to enhance the

precision of analyzing and identifying potato plant features. The structure of the MSFF module is shown in Figure 7.

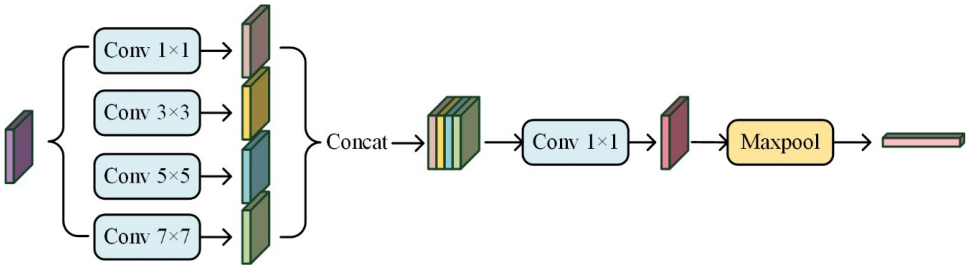


Figure 7. MSFF module structure.

The structure achieves multi-scale feature extraction by convolving the input image with filters of different sizes, including 1×1, 3×3, 5×5, and 7×7. These convolution kernels are each responsible for capturing details at different scales in the image. The feature maps of these four scales are then spliced in the channel dimension to form a feature tensor containing multi-scale information. Next, these features are fused using a 1×1 convolutional layer to adjust the number of output channels. Finally, a maximum pooling operation is applied to downsample the feature map, resulting in a compact and information-rich feature representation.

The MSFF module aims to enhance the model's comprehension of both details and global information in the image, yet this also results in an escalation of the network parameters. In the MSFF module, each convolutional layer contains a set of learnable convolutional kernels, which are the main source of parameter growth. Specifically, the MSFF module contains five convolutional layers, each of which introduces a certain number of parameters based on its convolutional kernel size, number of input channels (in_c), and number of output channels (out_c). For example, a convolutional layer will have one parameter if its convolutional kernel size is $k \times k$. This stacking of multiple convolutional layers significantly amplifies the total number of parameters in the network.

2.4. Residual Learning

Deep convolutional neural networks have achieved remarkable milestones in the field of computer vision. With an increase in the number of network layers, the extracted image features become more intricate, leading to enhanced training effectiveness. Nevertheless, deepening the network further can lead to the problem of network degradation. This issue manifests as the training set loss initially decreases, then plateaus or even increases as the network depth increases. To overcome this problem, He et al. [27] proposed the residual structure, which introduces shallow features by adding jump connections between layers.

The fundamental concept of residual learning is to alter the intricate mapping from input to output that conventional neural networks learn directly. Instead, the network is trained to learn a more straightforward "residual" mapping, which represents the discrepancy between the input and the desired output. In mathematical terms, if the function that our network wishes to learn is, then the residual learning approach is to have the network learn $F(x) = H(x) - x$. This allows the final output, y, to be obtained via the equation $y = F(x) + x$. The key to this transformation is that when learning $H(x)$ directly is difficult (especially when the network is deep), learning a relatively simple $F(x)$ becomes more feasible. The residual structure is shown in Figure 8.

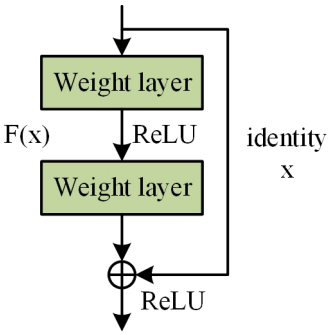


Figure 8. Residual structure.

In order to implement the aforementioned residual mapping idea, residual networks introduce a technique known as "skip connection". A skip connection is a part of the network where the input signal is permitted to bypass one or more layers and pass directly to the next layer, where it is added to the transformed signals of those layers to construct the final output. This approach not only simplifies the learning task but also ensures that the forward and backward propagation paths of the network are unobstructed, especially in deep networks, which helps to mitigate the problem of vanishing or exploding gradients.

The utilization of residual learning represents an efficacious methodology for enhancing the performance of networks in image recognition tasks. The strategy enables the network to concentrate on identifying subtle discrepancies between the inputs and the desired outputs by learning the residuals instead of the complete output mapping, a process that is more straightforward than directly mastering the global mapping model. As a result, residual networks achieve superior convergence and enhance the model's generalization performance on previously unseen data. The introduction of a residual learning mechanism effectively reduces the task of understanding complex functions to the learning of simpler residual functions, thereby reducing the learning burden and ensuring that each layer of the network works on refining existing features and incorporating new information rather than reconstructing all the features of the previous layer repeatedly. It is of particular importance to note that information can be freely passed across multiple layers through the shortcut connection architecture. This allows higher layers to directly utilize features from lower layers, which is crucial for retaining and transmitting critical information in tasks such as image recognition.

2.5. Overall Structure of the MSR-SwinT Potato Plant Identification Model

In order to address the challenge of complex trait recognition in potato plants, two improvements were made in this study based on the Swin Transformer model. Firstly, the multi-scale feature fusion module was employed in place of Patch Partition and Linear Embedding. Secondly, the residual learning strategy was introduced into the Swin Transformer Block. The improved model is named MSR-SwinT, and its overall structure is shown in Figure 9, and its detailed configuration is shown in Table 3.

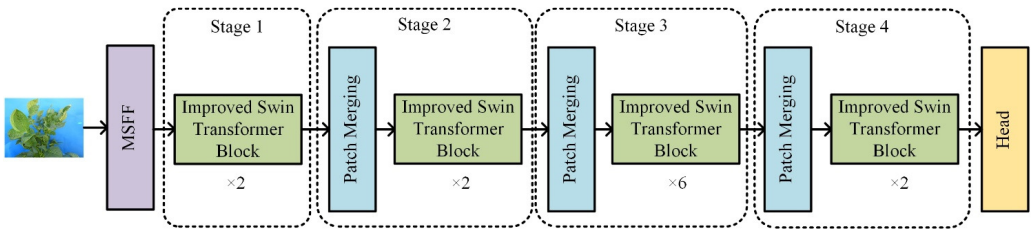


Figure 9. MSR-SwinT overall structure.

Table 3. Detailed configuration of MSR-SwinT.

Output Size	Module Name	Configuration	
56×56	MSFF	$\begin{bmatrix} \text{conv } 1 \times 1, \text{stride} = 2 \\ \text{conv } 3 \times 3, \text{stride} = 2, \text{padding} = 1 \\ \text{conv } 5 \times 5, \text{stride} = 2, \text{padding} = 2 \\ \text{conv } 7 \times 7, \text{stride} = 2, \text{padding} = 3 \end{bmatrix}$ conv 1×1, stride = 1 Maxpool, stride = 2	concat
	Improved Swin Transformer Block	$\begin{bmatrix} \text{window size } 7 \times 7 \\ \text{dim} = 96, \text{head} = 3 \end{bmatrix}$	$\times 2$
28×28	Patch Merging	Concat 2×2, 192-d, LayerNorm	
	Improved Swin Transformer Block	$\begin{bmatrix} \text{window size } 7 \times 7 \\ \text{dim} = 192, \text{head} = 6 \end{bmatrix}$	$\times 2$
14×14	Patch Merging	Concat 2×2, 384-d, LayerNorm	
	Improved Swin Transformer Block	$\begin{bmatrix} \text{window size } 7 \times 7 \\ \text{dim} = 384, \text{head} = 12 \end{bmatrix}$	$\times 2$
7×7	Patch Merging	Concat 2×2, 768-d, LayerNorm	
	Improved Swin Transformer Block	$\begin{bmatrix} \text{window size } 7 \times 7 \\ \text{dim} = 768, \text{head} = 24 \end{bmatrix}$	$\times 2$
1×1	Head	AdaptiveAvgPool, Flatten, Linear	

Potato plant images were first run through a designed multi-scale feature fusion module. The conventional patch partition method divides the image into fixed-size blocks, potentially overlooking global information and resulting in a decline in the model's performance when handling images of diverse scales. Conversely, the multi-scale feature fusion module can adeptly manage features of varying scales, enhancing the model's versatility. This module considers both local and global information, enabling the model to gain a deeper understanding of the image's structure and content. Hence, in the Swin Transformer model, the initial patch partition and the linear embedding layer of stage 1 are substituted with the multi-scale feature fusion module. This enhancement is designed to boost the model's performance and generalization capabilities.

The resultant multi-scale fused features are subsequently input into the enhanced Swin Transformer Block for further processing. In the task of plant image recognition, it is pivotal to thoroughly extract and leverage the features to enable the network to capture more valuable information. Convolutional neural networks are commonly employed in plant image recognition tasks. To enhance network performance and mitigate the issue of model degradation due to excessive network depth, residual connections are frequently employed in convolutional networks. In this chapter, the concept of residual learning is integrated into the Swin Transformer block, illustrated in Figure 10. Compared with the original Swin Transformer Block, this improved Swin Transformer Block retains its layering and moving window self-attention mechanism and further incorporates residual learning, which not only allows the network to expand to a deeper level but also ensures that the gradient can be passed directly to the depth of the network, effectively mitigating the problem of disappearing gradients. The design of the residual connectivity allows the original input signal to bypass the intermediate layer and reach the back part of the network directly, facilitating the smooth transfer of information through the network. In addition, this structure helps the network learn more complex feature representations, which in turn improves overall performance in recognition and classification tasks.

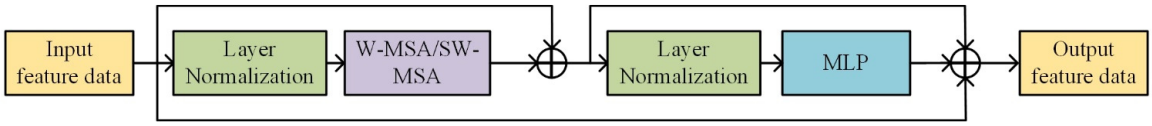


Figure 10. Improved Swin Transformer Block.

Next, the feature mapping processed by the enhanced Swin Transformer block undergoes the operation of the patch merging layer. This stage mainly involves aggregation and dimensionality reduction of local neighborhood image patches to construct feature representations at a higher level of abstraction, thereby reducing the spatial dimensionality of the model without losing the most important semantic information.

Finally, the high-level features formed after a series of the above processes are fed directly into the classification layer. Based on the previously accumulated multi-level and multi-scale feature information, the classification layer generates probability distributions of the attributes of each category of potato plants through the MLP head to arrive at the most probable trait categories as the final output.

In conclusion, the MSR-SwinT model builds a powerful tool for complex trait analysis of potato plants by cleverly integrating the advantages of multi-scale feature extraction, residual learning, and Swin Transformer, and this approach is expected to significantly improve the automation and intelligence of image-based analysis of potato.

2.6. Experimental Environment and Parameter Setting

The experiment was conducted in a computer environment running Windows 11, powered by a 13th Gen Intel(c) Core(TM) i7-13700 processor and an NVIDIA GeForce RTX 4090 GPU. Details of the experimental environment configuration are provided in Table 4.

Table 4. Experimental environment configuration.

Experimental Environment	Configuration Parameters
Operating system	Windows11
Processor	13th Gen Intel(c) Core(TM) i7-13700
GPU	NVIDIA GeForce RTX 4090
Code management software	Pycharm 2020.1.3
Programming languages	Python3.8
Deep learning frameworks	Pytorch1.12.1
GPU acceleration libraries	CUDA 11.6.134

The model was trained using the AdamW optimization algorithm with a weight decay of 5E-2. The Cross-Entropy Loss function was chosen for calculation. During model training, the batch size was set to 16, the epoch to 100, and after several iterations and debugging, the learning rate was eventually set to 0.00002.

2.7. Indicators for Model Evaluation

In the field of deep learning, model evaluation metrics are a central means of assessing model performance, helping to understand and compare the performance of different models in depth while playing a crucial role in the model selection and optimization phases. Since different tasks have their own unique training objectives, the metrics used to evaluate model performance vary and need to be selected for specific tasks. For multi-classification tasks, it is difficult for a single evaluation metric to fully reflect model performance. Therefore, multiple evaluation metrics need to be used to evaluate the model performance fully.

The main evaluation metrics used in this paper to evaluate the model are accuracy, precision, recall, specificity, F1 value, and confusion matrix. Accuracy is a measure of the percentage of correctly classified samples by the model out of the total test set. It is the most straightforward evaluation metric, yet it may overlook the impact of category

imbalance. Precision is the percentage of all samples predicted by the model to be positive cases that are actually positive cases. Recall is the proportion of all samples that are actually positive cases that are correctly predicted as positive cases by the model. Specificity is the proportion of negative cases correctly predicted by the model out of all samples that are actually negative cases. F1 is a combination of precision and recall, balanced by the reconciled mean of precision and recall. The formulae for each indicator are shown in Equations (4)–(8).

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \tag{4}$$

$$\text{Precision} = \frac{TP}{TP + FP} \times 100\% \tag{5}$$

$$\text{Recall} = \frac{TP}{TP + FN} \times 100\% \tag{6}$$

$$\text{Specificity} = \frac{TN}{TP + FN} \times 100\% \tag{7}$$

$$\text{F1} = \frac{2 \times Precision \times Recall}{Precision + Recall} \times 100\% \tag{8}$$

where TP indicates that the model correctly identifies samples belonging to the positive category; FP represents instances where the model misclassifies samples in the negative category as positive; TN denotes a situation where the model successfully identifies samples belonging to the negative category; FN signifies that the model fails to identify samples that are, in fact, in the positive category and incorrectly classifies them as negative.

The confusion matrix [28] is a fundamental instrument for assessing the efficacy and precision of a model in classification tasks. It offers a comprehensive and readily comprehensible representation of the evaluation. By comparing the actual classification with the model predictions, the confusion matrix can classify the samples into four basic categories with a high degree of accuracy. The four categories of classification are true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). The confusion matrix provides a systematic outline of these basic counts, which directly reflect the accuracy and error distribution of the classification model in the judgment of each category. It is, therefore, an integral part of measuring and understanding model performance.

3. Experimental Results and Discussion

3.1. Analysis of Ablation Experiment Results

In order to verify the effectiveness of the improved model MSR-SwinT in the potato plant image classification task, the data_150 dataset is used as the experimental data, and the following experimental scheme is adopted: (a) using only the Swin Transformer model; (b) adding a multi-scale feature fusion module on top of the Swin Transformer model, denoted as Multi-scale-SwinT; (c) introducing residual learning in the Swin Transformer block, denoted as Res-SwinT; (d) using the improved method in both b and c, i.e., the improved model MSR-SwinT. The experimental results of the four removal schemes are shown in Table 5.

As can be seen from Table 5, Scheme b after integrating the multi-scale feature fusion module into the Swin Transformer model, the test set accuracy is increased by 2.88% compared with the original Swin Transformer model, and this result indicates that the multi-scale feature fusion module can effectively enhance the extraction of potato plant image features. In scheme c, the residual learning mechanism is introduced into the Swin Transformer block, and the test set accuracy of the model is increased by 1.39% compared

with the Swin Transformer model, which indicates the positive effect of residual learning in improving the model performance, which helps the model to better learn and retain the important feature information in the image. The improved model MSR-SwinT, which also uses scenarios b and c, achieved an accuracy of 94.64% on the test set, which is an improvement of 3.02% over the original Swin Transformer model, clearly demonstrating the superior performance of the improved model on the potato plant image classification task.

Table 5. Ablation experiments.

Model	Training Set Accuracy/%	Training Set Loss	Validation Set Accuracy/%	Validation Set Loss	Test Set Accuracy/%
Swin Transformer	92.51	0.23	92.27	0.23	91.62
Multi-scale-SwinT	95.09	0.15	94.91	0.15	94.24
Res-SwinT	93.54	0.20	92.98	0.21	92.75
MSR-SwinT	95.72	0.14	94.98	0.15	94.64

3.2. Analysis of the Impact of Learning Rate on the Model

Four different learning rate values were chosen in this paper: 0.0001, 0.00005, 0.00002, and 0.00001 were selected and experiments were conducted on the data_150 dataset. The training accuracy curves obtained from the experimental results are shown in Figure 11.

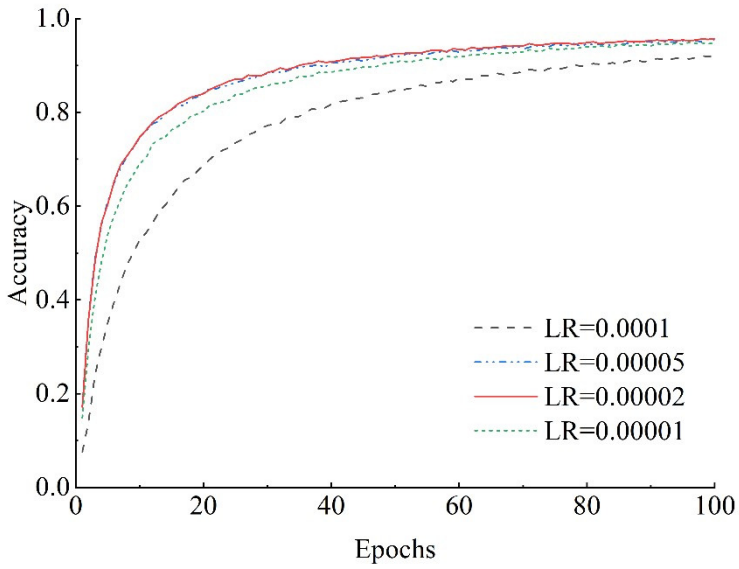


Figure 11. Training accuracy curves for different learning rates.

As can be seen from Figure 11, the model converges most slowly when the learning rate is 0.0001, and only begins to converge gradually after about 90 training iteration cycles. The other learning rates, on the other hand, begin to show a convergence trend after about 70 iteration cycles. Comparing the training results with a learning rate of 0.00005 and a learning rate of 0.00002, it was found that although there was not much difference in recognition accuracy between the two, the model with a learning rate of 0.00002 performed much better in terms of both convergence speed and final accuracy. All in all, it was found that setting the learning rate to 0.00002 was the best choice for the network model in this chapter.

3.3. Model Performance Evaluation

In order to conduct a comprehensive and in-depth analysis of the performance of the improved MSR-SwinT model, we examined its performance on the data_150 test set. The results are presented in Table 6. The results of the analysis demonstrate that the MSR-SwinT

model exhibits the highest accuracy in identifying the variety LS-19, with all 89 samples correctly predicted. Furthermore, the model demonstrated a high degree of accuracy in identifying the varieties L11133-2H, LS-25, and Ly1503-9. Only one sample in each variety was incorrectly predicted as the other varieties. However, the model demonstrated suboptimal performance in identifying the varieties L11120-8, Ly1602-3, and Ly1702-3, with 10 or 11 samples in each of these varieties being incorrectly predicted, respectively. In general, the MSR-SwinT model demonstrated satisfactory performance in the majority of variables, with precision, recall, and F1 values exceeding 90%. However, some challenges in identification persisted for certain varieties.

Table 6. MSR-SwinT recognition results on dataset data_150.

Class	TP	FP	FN	TN	Precision/%	Recall/%	F1 Score/%
L11120-8	82	3	10	2648	96.47	89.13	92.65
L11133-2H	90	1	1	2651	98.90	98.90	98.90
L11146-2	86	5	2	2650	94.51	97.73	96.09
L12116-19	88	8	3	2644	91.67	96.70	94.12
L13104-12	85	4	7	2647	95.51	92.39	93.92
L14140-5	85	1	7	2650	98.84	92.39	95.51
LS-12	88	6	5	2644	93.62	94.62	94.12
LS-17	84	2	2	2655	97.67	97.67	97.67
LS-19	89	1	0	2653	98.89	100	99.44
LS-20	85	3	3	2652	96.59	96.59	96.59
LS-23	82	0	6	2655	100	93.18	96.47
LS-25	86	8	1	2648	91.49	98.85	95.03
LS-26	85	2	2	2654	97.70	97.70	97.70
Ly1503-10	81	7	8	2647	92.05	91.01	91.53
Ly1503-9	91	13	1	2638	87.50	98.91	92.86
Ly1517-8	90	2	2	2649	97.83	97.83	97.83
Ly1520-9	88	4	4	2647	95.65	95.65	95.65
Ly1521-2	87	5	5	2646	94.57	94.57	94.57
Ly1544-2	91	1	5	2646	98.91	94.79	96.81
Ly1602-3	82	7	10	2644	92.13	89.13	90.61
Ly1615-3	85	8	8	2642	91.40	91.40	91.40
Ly1619-3	90	3	3	2647	96.77	96.77	96.77
Ly1624-1	86	7	6	2644	92.47	93.48	92.97
Ly1624-2	87	7	2	2647	92.55	97.75	95.08
Ly1628-1	87	6	5	2645	93.55	94.57	94.06
Ly1702-3	82	8	11	2642	91.11	88.17	89.62
Ly1722-1	88	3	5	2647	96.70	94.62	95.65
Ly1727-3	93	13	9	2628	87.74	91.18	89.43
Ly1733-16	85	2	8	2648	97.70	91.40	94.45
Ly1819-8	88	7	6	2642	92.63	93.62	93.12

3.4. Model Visualisation

In order to visualize the capability of the MSR-SwinT model on the potato plant identification task, the model was visualized and analyzed using feature maps and heat maps, respectively, and the results are shown in Figure 12. The feature maps of a single channel were selected for analysis in this process to understand the feature extraction mechanism within the model more finely. The Grad-CAM [29] was employed to generate heat maps, which highlighted key areas within the modeling decision-making process, thus further enhancing the visual analysis. The feature map illustrates that as the network layers deepen, the features extracted by the model evolve from low-level basic information, such as edges and textures, to high-level structural and semantic features. This reflects the continuous abstraction and generalization of the input information during the model learning process. The heat map demonstrates that the MSR-SwinT model is capable of effectively highlighting the salient features of the plant while simultaneously reducing the attention paid to irrelevant background factors.

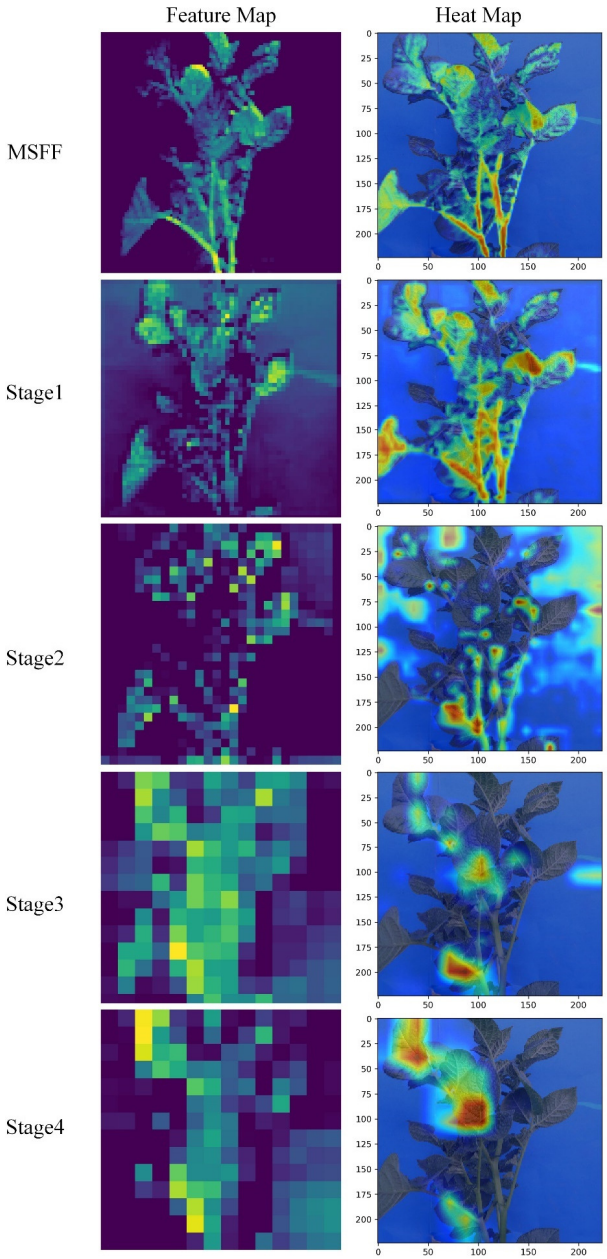
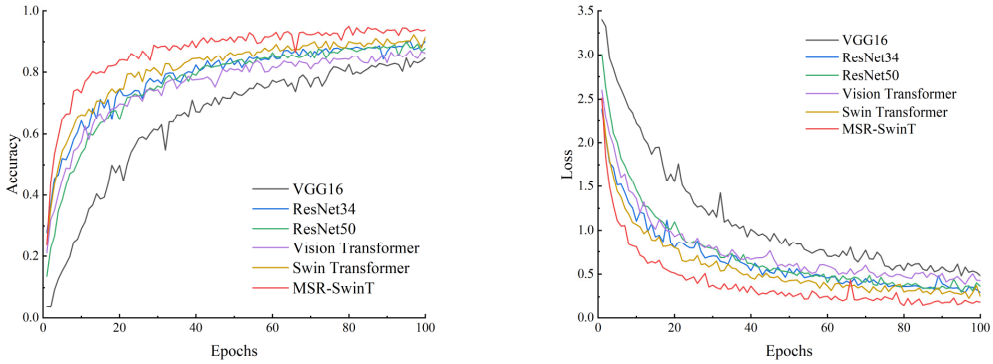


Figure 12. MSR-SwinT model visualization results.

3.5. Comparative Experimental Analysis of Different Models

In order to further analyze the performance of the improved model, the MSR-SwinT model is compared with the commonly used convolutional neural network models (including VGG16 [30], ResNet34, ResNet50) and Transformer models (Vision Transformer [31] and Swin Transformer) in the field of image recognition in the comparison experiments on the data_150 dataset, and the accuracy and loss curves of the validation set obtained are shown in Figure 13. As can be seen in Figure 13, the improved MSR-SwinT model shows faster convergence than the other models during training and reaches 90% accuracy after about 40 epochs, and at the same time, its loss value is significantly lower than other models. The VGG16 model has the slowest convergence speed and the lowest accuracy of all the models. Comparing the ResNet34 and ResNet50 models, the ResNet34 model converges faster than ResNet50 and has a slightly higher final accuracy than ResNet50. The reason for this is that ResNet50 can cause gradient explosion, whereas the shallower network, ResNet34, may be better at avoiding the gradient explosion problem.



(**a**) (**b**)

Figure 13. Accuracy curve and loss curve of validation set of different models. (**a**) Accuracy curve, (**b**) Loss curve.

The best weights obtained from the validation set are applied to the test set, and the experimental results obtained are shown in Table 7. As can be seen from Table 7, the improved model MSR-SwinT outperforms the other four models in terms of accuracy, precision, recall, specificity, and F1 value, with indicators of 94.64%, 94.77%, 94.69%, 99.82%, and 94.69% respectively. This result indicates that the improved MSR-SwinT model shows high performance and effectiveness in detecting potato plants, and its excellent performance in key indicators such as accuracy confirms the good applicability of the model in specific application scenarios. In particular, the model can accurately identify and distinguish potato plants, providing a reliable tool for relevant research and practice.

Table 7. Results of different model test sets.

Model	Accuracy/%	Precision/%	Recall/%	Specificity/%	F1/%
VGG16	84.07	84.89	84.12	99.45	84.13
ResNet34	90.05	90.30	90.07	99.66	90.05
ResNet50	90.01	90.27	90.02	99.65	89.97
Vision Transformer	86.66	87.15	86.62	99.54	86.72
Swin Transformer	91.62	91.78	91.64	99.71	91.63
MSR-SwinT	94.64	94.77	94.69	99.82	94.69

The confusion matrix for each model on the data set data_150 is presented in Figure 14. The depth of the color of the diagonal line in the confusion matrix represents the recognition accuracy of each category, with darker colors indicating a higher correct recognition rate for that category. As illustrated in Figure 14, the confusion matrix of the MSR-SwinT model exhibits a darker diagonal color compared to the other models. The MSR-SwinT model demonstrated particularly strong performance in identifying the LS-19 variety, correctly predicting all 89 samples. For the varieties L11133-2H, LS-25, and Ly1503-9, the model also performed well, with only one sample misclassified for each of these varieties. This observation provides compelling evidence that the MSR-SwinT model demonstrates superior performance in recognizing potato plant species, exhibiting the highest recognition accuracy for each species.

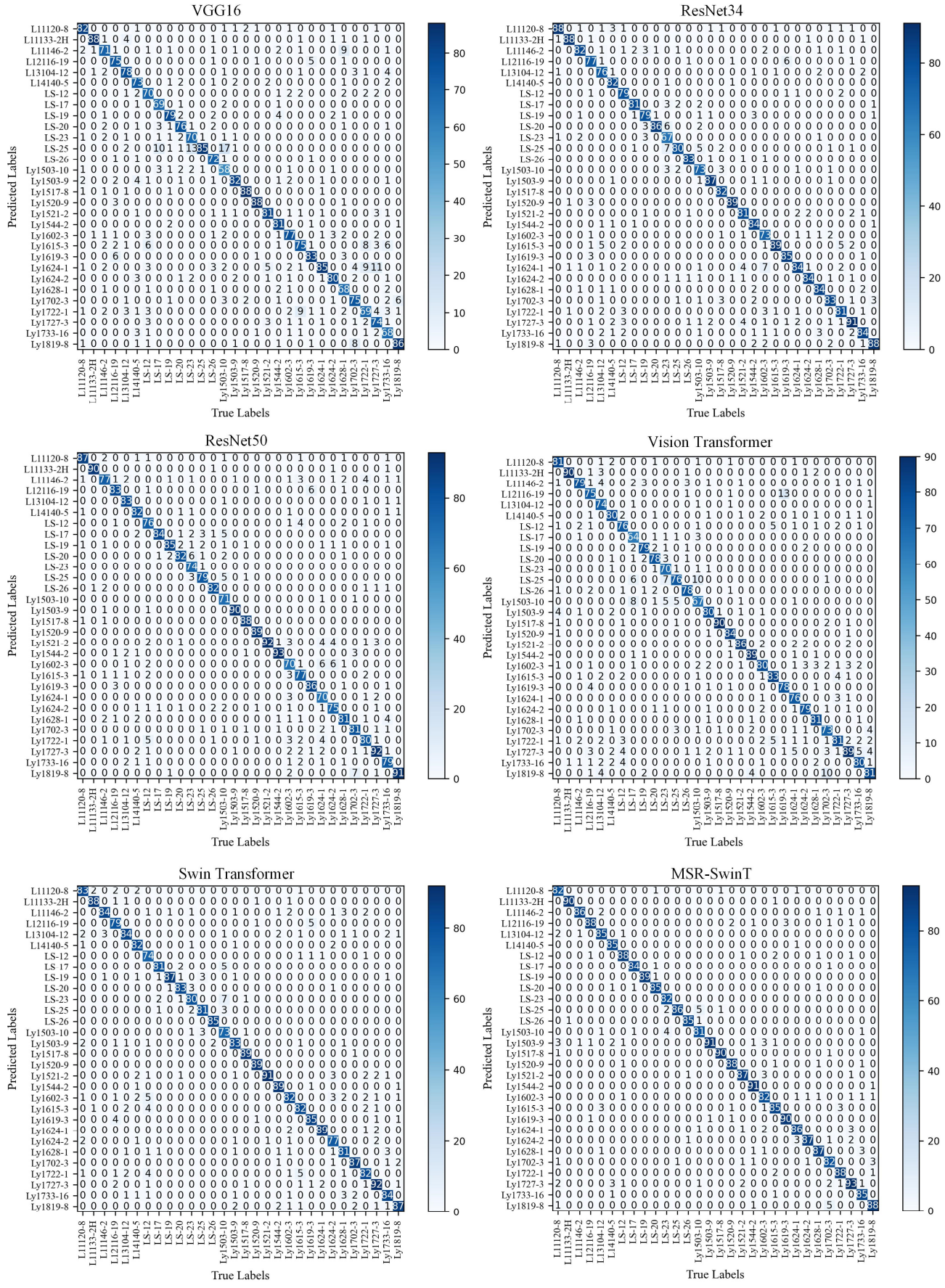


Figure 14. Confusion matrix for different models.

3.6. Analysis of Experimental Results with Different Sample Sizes

In order to investigate the effect of different sample sizes on the model performance, four new data sets were constructed from the original data set data_150 by randomly selecting a different number of samples for each variety. Specifically, 100, 70, 50, and 40 plants were randomly selected for each variety to create four data sets of different sizes. Such an experimental design helps to analyze the variation of the model performance under different sample size conditions in order to assess the stability and robustness of the model. The accuracy performance of the VGG16, ResNet34, ResNet50, Vision Transformer, Swin Transformer, and MSR-SwinT models on the validation sets with sample sizes of 150, 100, 70, 50, and 40 stems is shown in Table 8.

Table 8. Validation set accuracy for different models with different sample sizes.

Model	Accuracy/%				
	150 Plants	**100 Plants**	**70 Plants**	**50 Plants**	**40 Plants**
VGG16	84.58	77.83	76.11	71.56	61.81
ResNet34	90.26	90.06	88.89	87.89	85.69
ResNet50	90.04	88.61	85.83	84.45	83.75
Vision Transformer	87.14	86.22	85.79	84.78	81.94
Swin Transformer	92.27	90.61	89.84	87.78	85.97
MSR-SwinT	94.98	94.12	93.91	93.78	90.28

From Table 8, it can be clearly seen that the improved MSR-SwinT model shows better recognition effects than other models on datasets of different sizes. By analyzing several datasets with different sample sizes, it can be seen that the recognition performance of each model gradually improves as the sample size increases. The reason for this phenomenon is that more data provides the models with richer and more diverse feature information, which not only helps the models to capture more subtle category differences but also enables the models to demonstrate higher accuracy and stronger generalization ability in the recognition task. Therefore, this finding suggests that with sufficient data, each model can demonstrate significant advantages and that in real-world application scenarios, increasing the amount of data collected is likely to further improve the performance of the models.

3.7. Analysis of Experimental Results on Different Data Sets

The Swin Transformer model and the MSR-SwinT model are experimented on different data sets data_100, data_70, data_50, and data_40, and the experimental results obtained are shown in Table 9.

Table 9. Experimental results for different data sets.

Model	Accuracy/%			
	Data_100	**Data_70**	**Data_50**	**Data_40**
Swin Transformer	89.28	85.60	79.76	83.24
MSR-SwinT	93.56	92.90	87.57	90.46

The data sets data_100, data_70, data_50, and data_40, which have undergone the data enhancement process, are now named aug_data_100, aug_data_70, aug_data_50, and aug_data_40, respectively. The Swin Transformer model with the MSR-SwinT model is then applied to the augmented dataset, and the resulting experimental results are presented in Table 10.

Table 10. Experimental results of data enhancement for different datasets.

Model	Accuracy/%			
	Data_100	**Data_70**	**Data_50**	**Data_40**
Swin Transformer	94.27	93.28	89.90	91.42
MSR-SwinT	97.01	96.42	94.73	95.43

It can be observed from Table 9 that the accuracy of the improved MSR-SwinT model is generally higher than that of the Swin Transformer model on different datasets, which proves the effectiveness of the improvement measures taken in this chapter and indicates that the MSR-SwinT model has a better performance in recognizing potato plant images. Meanwhile, it can be observed in Table 10 that the accuracy of both the Swin Transformer model and the MSR-SwinT model is improved after data expansion, indicating that data expansion is an effective means to improve the efficiency of model recognition.

Tables 9 and 10 show not only the performance of the MSR-SwinT model on different data sets but also an interesting phenomenon: although the data volume of data_40 is smaller than that of data_50, its accuracy rate is relatively high. By carefully comparing the original images of these two datasets, it was found that the images in the data_50 dataset contained more background and other distracting factors, which may have negatively affected the model's recognition ability, leading to its relatively low accuracy. This further emphasizes that in an image dataset, in addition to the number of samples, the quality and relevance of the samples are also key factors affecting the performance of the model. Therefore, careful screening and preprocessing of images to minimize the influence of irrelevant factors during the data preparation stage is essential to improve model accuracy.

3.8. Discussion

The application of deep learning technology in the field of potato variety identification can significantly improve identification accuracy, thus playing a positive role in ensuring potato germplasm resource management and food safety. The convolutional neural network has strong local image feature extraction ability, can achieve end-to-end learning and inference in the whole recognition process, and effectively extracts key local features from the input image so as to achieve accurate recognition tasks. The unique feature of Visual Transformer is the use of a self-attention mechanism, which can effectively capture the interdependence between any two points in the image, which is very important for identifying the key features of varieties, especially in processing global contextual information and discovering potential fine-grained differences.

In this work, potato plant datasets with different sample sizes were collected and constructed, containing a total of 69 potato varieties. In order to effectively identify potato plants, the Swin Transformer model was improved. The multi-scale feature fusion module was used to replace the patch partition of the Swin Transformer model and the linear embedding of stage 1, which not only effectively fused the features of different scales but also improved the generalization ability of the model; at the same time, residual learning was introduced into the Swin Transformer block mechanism, which helps the smooth flow of information and enhances feature transfer, thus preserving more complex features in the image. The recognition accuracy of the improved MSR-SwinT model is improved by 3.02 percentage points over the original Swin Transformer model.

Different learning rates significantly affect the training results of the model. Adjusting the learning rate can affect the convergence speed, stability, and final performance of the model. A lower learning rate may make the model training more stable, but it may lead to a slower convergence speed, while a higher learning rate may speed up the convergence speed, but it may increase the fluctuation during the training process or even lead to the

model failing to converge. Therefore, choosing an appropriate learning rate is crucial for model optimization. By setting different learning rates for comparison experiments, we finally decided to set the learning rate to 0.00002.

The amount of data is also a key factor that significantly affects model performance. Through a series of experiments with different sample sizes, a clear trend can be observed: as the amount of training data increases, the recognition performance of the model shows a positive improvement. In other words, the more sufficient the amount of data, the better the model performs in the recognition task.

In addition, the Swin Transformer base model and the improved MSR-SwinT model are selected for experimental validation on several different datasets in this paper. In order to further validate the effect of data volume on model performance, appropriate data expansion methods are used on different datasets. The experimental results show that the improved MSR-SwinT model embodies a significant performance improvement on the potato plant identification task, confirming the effectiveness of the improvement measures. Furthermore, the extension of the dataset helps to improve the detection performance of the model. Additionally, the experiments reveal that the quality of the data images significantly impacts the model's performance and stands as one of the most crucial factors influencing the model's recognition accuracy.

4. Conclusions

In response to the lack of research on potato variety classification and considering the characteristics of potato plants with vigorous growth, dense foliage, and their stems usually growing upright or creeping and spreading, a potato plant recognition model, MSR-SwinT, was proposed. The Swin Transformer model was improved by using a multi-scale feature fusion module and a residual learning mechanism, and the improved MSR-SwinT model performed well on the task of potato plant classification with an accuracy of 94.64% in the test set. The improvement in classification performance of MSR-SwinT was further confirmed by comparing the confusion matrices of the Swin Transformer and MSR-SwinT models.

To further assess the performance of MSR-SwinT, comparative experiments were conducted against VGG16, ResNet34, ResNet50, and Vision Transformer models. The results indicate that the recognition accuracy of the MSR-SwinT model surpasses these models by 10.57%, 4.59%, 4.63%, and 7.89%, respectively. This demonstrates the superior performance of the MSR-SwinT model in the potato plant recognition task.

The MSR-SwinT model exhibits exemplary visual information processing capabilities, as evidenced by its integrated multi-scale feature module, residual learning method, and Swin Transformer technology. The model enhances the sensitivity to various types of dimensional features through multi-scale feature processing, thus deepening the comprehensive understanding of the local subtle features and the overall layout of the image and scene. The integration of the residual learning mechanism effectively addresses the pervasive issue of gradient disappearance in deep network training, thereby enhancing the model's capacity to learn sophisticated feature representations, expediting the training process, and improving the model's performance. The Swin Transformer's core is its innovative self-attention computation system and dynamic moving window strategy, which work in concert to achieve a notable enhancement in computational efficiency. This is achieved while preserving the expressive power of the Transformer architecture.

At present, there is a paucity of research literature on the identification of potato plant varieties. This study provides a foundation for subsequent research on potato variety identification and technical support for potato variety safety. Nevertheless, although MSR-SwinT is effective in enhancing processing power and efficiency, its intricate structural design

and elevated resource requirements are significant, particularly when hardware resources are constrained. Furthermore, if the objective is to deploy it on mobile, these factors become significant challenges that must be addressed. In the future, model compression, quantization, pruning, and other techniques may be employed to reduce the model size and optimize the computational process, thereby achieving the best performance under conditions of limited resources.

In the future, we will focus on enhancing the model's performance and expanding its applications in multiple aspects. On the one hand, we will extensively collect potato data from different regions and seasons and construct time series to enrich the diversity of samples. Meanwhile, we will compare new models that emerged in the last two years to optimize the MSR—SwinT. On the other hand, we will actively explore its practical applications in potato cultivation, such as integrating it into an automated identification system in the field to achieve precise planting management. We will also investigate its use in other links to the industrial chain. In addition, through targeted adjustments, the model will be adapted to various crop varieties and expanded to agricultural tasks such as pest and disease detection and ecological environment monitoring, providing comprehensive technical support for the sustainable development of agriculture.

Author Contributions: Conceptualization, C.L.; methodology, X.X., C.L. and J.H.; software, X.X.; validation, C.L., J.H. and Q.F.; formal analysis, C.L.; investigation, X.X., C.L. and J.H.; resources, E.Q. and Y.Q.; data curation, X.X. and B.M.; writing—original draft preparation, X.X.; writing—review and editing, X.X. and C.L.; visualization, X.X.; supervision, C.L. and J.H.; project administration, C.L.; funding acquisition, C.L. and J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Fund Project (Grant 32360437), the Industrial Support and Guidance Project of Universities in Gansu Province of China (Grant 2021CYZC-57), and the Innovation Fund Project of Colleges and Universities in Gansu of China (Grant 2021A-056).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available because they are still being analyzed.

Acknowledgments: We thank the editor and reviewers for their helpful suggestions to improve the quality of this manuscript.

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

References

1. Yang, Y.; Li, G.; Cheng, C. Optimisation of potato planting layout and restructuring of positioning: An elaboration of social embeddedness theory. *Rural Econ.* **2023**, *03*, 48–57.
2. Wang, X.; Chen, M. Retrospect and prospect of potato development. *Issues Agric. Econ.* **2020**, 123–130. [CrossRef]
3. Zhong, Y. Uses of potato and its development prospect. *Rural Sci. Technol.* **2016**, 39–41. [CrossRef]
4. Naznin, F.; Haque, M.A.; Rahman, M.H.S.; Rahman, M.H.A.; Mehedi, M.N.H. Molecular Diversity Analysis of Indigenous Potato Cultivars Using RAPD Markers. *Asian J. Biochem. Genet. Mol. Biol.* **2020**, *6*, 43–48. [CrossRef]
5. Yin, X.; Zhang, X. Application of AFLP molecular markers in genetic breeding of eggplant and fruit vegetables. *Chin. Agric. Sci. Bull.* **2009**, *25*, 25–31.
6. Wang, X.; Hui, Z.; Xu, J.; Duan, S.; Bian, C.; Jin, L.; Li, G. Histo-cytological study on potato potato shape development. *China Veg.* **2020**, 67–73. [CrossRef]
7. Hu, H. *Fundamentals of Machine Learning*; Electronic Industry Press: Beijing, China, 2019; p. 384.
8. Zhang, H.; Zhang, Y.; Wang, L.; Hu, Z.; Zhou, W.; Tsang, P.W.; Cao, D.; Poon, T.-C. Study of Image Classification Accuracy with Fourier Ptychography. *Appl. Sci.* **2021**, *11*, 4500. [CrossRef]
9. Zhang, Z.; Ji, Y.; Man, W. Improved random forest algorithm for image classification applications. *Comput. Syst. Appl.* **2018**, *27*, 193–198.

10. Fu, L.; Zhang, M.; Liu, Z.; Li, H.; He, T. Adaptive hybrid feature extraction for leaf image classification by support vector machine. *J. Converg. Inf. Technol.* **2012**, *7*, 160–166.
11. Xu, j.; Wang, J.; Xu, X.; Ju, S. Rice fertility image recognition based on RAdam convolutional neural network. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 143–150.
12. Li, J.; Xu, F.; Song, S.; Qi, J. A maize seed variety identification method based on improving deep residual convolutional network. *Front. Plant Sci.* **2024**, *15*, 1382715. [CrossRef] [PubMed]
13. Taheri-Garavand, A.; Nasiri, A.; Fanourakis, D.; Fatahi, S.; Omid, M.; Nikoloudakis, N. Automated In Situ Seed Variety Identification via Deep Learning: A Case Study in Chickpea. *Plants* **2021**, *10*, 1406. [CrossRef] [PubMed]
14. Liu, C.; Han, J.; Chen, B.; Mao, J.; Xue, Z.; Li, S. A Novel Identification Method for Apple (*Malus domestica* Borkh.) Cultivars Based on a Deep Convolutional Neural Network with Leaf Image Input. *Symmetry* **2020**, *12*, 217. [CrossRef]
15. Gao, J.; Liu, C.; Han, J.; Lu, Q.; Wang, H.; Zhang, J.; Bai, X.; Luo, J. Identification Method of Wheat Cultivars by Using a Convolutional Neural Network Combined with Images of Multiple Growth Periods of Wheat. *Symmetry* **2021**, *13*, 2012. [CrossRef]
16. Laabassi, K.; Belarbi, M.A.; Mahmoudi, S.; Mahmoudi, S.A.; Ferhat, K. Wheat varieties identification based on a deep learning approach. *J. Saudi Soc. Agric. Sci.* **2021**, *20*, 281–289. [CrossRef]
17. Guo, Y. *Research and Application of Potato Variety Classification Based on Convolutional Neural Network*; Gansu Agricultural University: Lanzhou, China, 2023.
18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
19. Liu, Y.; Liu, H.; Shi, T.; Ouyang, J.; Huang, H.; Xie, T. An optimised Swin Transformer tomato leaf disease identification method. *J. China Agric. Univ.* **2023**, *28*, 80–90.
20. Fu, L.; Huang, H.; Wang, H.; Huang, S.; Chen, D. Classification of maize growing period based on Swin Transformer model. *Trans. Chin. Soc. Agric. Eng.* **2022**, *38*, 191–200.
21. Wang, F.; Rao, Y.; Luo, Q.; Jin, X.; Jiang, Z.; Zhang, W.; Li, S. Practical cucumber leaf disease recognition using improved Swin Transformer and small sample size. *Comput. Electron. Agric.* **2022**, *199*, 107163. [CrossRef]
22. Reedha, R.; Dericquebourg, E.; Canals, R.; Hafiane, A. Transformer Neural Network for Weed and Crop Classification of High Resolution UAV Images. *Remote Sens.* **2022**, *14*, 592. [CrossRef]
23. Borhani, Y.; Khoramdel, J.; Najafi, E. A deep learning based approach for automated plant disease classification using vision transformer. *Sci. Rep.* **2022**, *12*, 11554. [CrossRef] [PubMed]
24. Luo, J. *Research on Variety Recognition and Classification of Potato Images Based on Deep Learning*; Gansu Agricultural University: Lanzhou, China, 2023.
25. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 10012–10022.
26. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
27. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
28. Chen, J.; Han, J.; Liu, C.; Wang, Y.; Shen, H.; Li, L. A Deep-Learning Method for the Classification of Apple Varieties via Leaf Images from Different Growth Periods in Natural Environment. *Symmetry* **2022**, *14*, 1671. [CrossRef]
29. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
30. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556. [CrossRef]
31. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.

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