Lychee cultivar fine-grained image classification method based on improved ResNet-34 residual network

Yiming Xiao,[1,3] Jianhua Wang,[1,2,3] Hongyi Xiong,[1,3] Fangjun Xiao,[1,3] Renhuan Huang,[1,3] Licong Hong,[1,3] Bofei Wu,[1,3] Jinfeng Zhou,[1,3] Yongbin Long,[1,2,3] Yubin Lan[1,2,3]

[1]College of Electronic Engineering (college of Artificial Intelligence); [2]Guangdong Laboratory for Lingnan Modern Agriculture; [3]National Center for International collaboration Research on precision Agricultural Aviation Pesticides Spraying Technology, South China Agricultural University, Guangzhou, China

Abstract

Lychee, a key economic crop in southern China, has numerous similar-looking varieties. Classifying these can aid farmers in understanding each variety's growth and market demand, enhancing agricultural efficiency. However, existing classification techniques are subjective, complex, and costly. This paper proposes a lychee classification method using an improved ResNet-34 residual network for six common varieties. We enhance the CBAM attention mechanism by replacing the large receptive field in the SAM module with a smaller one. Attention mechanisms are added at key network stages, focusing on crucial image information. Transfer learning is employed to apply ImageNet-trained model weights to this task. Test set evaluations demonstrate that our improved ResNet-34 network surpasses the original, achieving a recognition accuracy of 95.8442%, a 5.58 percentage point improvement.

Introduction

Lychee (scientific name: *Litchi Chinensis Sonn*) originated in China (Chang, 1961) and is an important economic crop in the southern regions of China. Lychee is widely distributed in southwestern, southern, and southeastern China, with a particular concentration in the Fujian and Guangdong provinces (Mitra and Pathak, 2008; Jiang *et al.*, 2021).

According to incomplete statistics, there are approximately more than 200 recorded varieties of lychee worldwide (Wu, 1998; Menzel *et al.*, 2005; Chang *et al.*, 2017). The large number of varieties and their high similarity pose many challenges to lychee production and research (Khurshid *et al.*, 2004). Firstly, it misleads consumers and affects market competition. For example, in the Guangzhou market, there are only around 20 varieties of lychee available for purchase, yet visually similar varieties can have significant price differences, making it difficult for consumers to distinguish them. Secondly, it jeopardizes the conservation of germplasm resources. Correctly identifying varieties is a crucial step in preserving the diverse genetic resources of lychee, and the confusion of varieties diminishes their utility (Zhang *et al.*, 2013; Yao *et al.*, 2021). Thirdly, there is a problem of chaotic naming of lychee varieties within different regions (Aradhya *et al.*, 1995), where the same variety may be assigned different names, and different varieties may share the same or similar names.

Various scholars from different fields have endeavored to classify lychee varieties. Liu *et al.* (2015) conducted molecular-level identification of lychee varieties and their genetic relationships using single nucleotide polymorphism (SNP) markers. Similarly, Madhou and colleagues (2013) differentiated between different lychee germplasms using microsatellite markers. While these molecular and genetic identification methods boast high accuracy, their operational complexity and time-consuming nature pose significant challenges. In the realm of machine vision-based lychee variety classification, Liu *et al.* (2014) explored the utility of hyperspectral imaging (HSI) technology and multivariate classification in lychee identification. Their work established correlations between reflectance spectra and lychee varieties, achieving a recognition rate of 87.81% using an SVM model. However, the

reliance on hyperspectral imaging entails the need for specialized equipment, escalating research costs and experimental complexities. Moreover, the technology's classification accuracy and reliability are susceptible to environmental influences. In a separate study, a Japanese scholar (Osako *et al.*, 2020) employed deep learning technology to discern lychee varieties. Their findings indicated that an enhanced VGG-16 network achieved a remarkable 99% accuracy in identifying four lychee varieties. Nevertheless, the study's narrow dataset background and limited lychee variety selection (four) may restrict the model's generalization capability, limiting its applicability and practicality in real-world scenarios.

With the continuous development of artificial intelligence technology, the field of fine-grained image classification in agriculture has also begun to widely adopt AI algorithms (Zhao *et al.*, 2018; Wang *et al.*, 2021; Shaikh *et al.*, 2022; Akkem *et al.*, 2023). In this trend, the widespread application of ResNet has become a prominent feature. This can be attributed to its introduction of residual blocks, which allow the model to learn residual information through skip connections, helping to address the problem of vanishing gradients (Akkem *et al.*, 2023). This has made it possible to design deeper neural networks, overcoming the difficulties faced by traditional deep networks during training (Yu *et al.*, 2022b; Akkem *et al.*, 2023). The advantages of ResNet are not only reflected in better training effects and generalization performance, but also in its simple and clear structure, which is easy to understand and interpret (Yu *et al.*, 2022a), making it the preferred backbone network for image tasks in various fields (Yu *et al.*, 2022b). A continuous emergence of ResNet-based variant networks, introducing improvements and innovations based on it, further adapts to different task requirements, expanding the possibilities for its application range. For example: Stephen *et al.* (2023) designed a self-attention ResNet network architecture for classifying rice leaf diseases, Wu *et al.* (2023) incorporated the SE attention mechanism and Ranger optimizer into the ResNet-50 network to improve its capability to recognize chicken gender, and Wang *et al.* (2022) proposed the S-ResNet model for identifying insects in crops, adding a residual structure branch to the base ResNet. Sennan *et al.* (2022) proposed an improved residual network structure for the identification of spinach varieties. Additionally, deep learning technology has brought significant advancements to multitask medical image analysis (Jiang *et al.*, 2023; Zhao *et al.*, 2023).

Based on the above, in order to better solve the lychee variety classification problem, this study collected 3799 images of six lychee varieties. To address the problem of overfitting in deep neural networks due to the similarity of features between certain lychee varieties, this study improved the ResNet-34 backbone network to improve the accuracy of lychee variety classification. Specifically, the convolutional block attention module (CBAM) attention mechanism was introduced, and the spatial attention module (SAM) in the mechanism was improved. The improved attention mechanism module was added after Global Max_Pooling, before entering Global Avg_Pooling, and after stage 3 of ResNet-34 to help the model more accurately locate and identify the key information in the image. To further improve the model's generalization ability, transfer learning technology was used, and the weights of the ImageNet classification task were transferred to this study's task, and data augmentation was applied to the input data. The CBAM attention mechanism can adaptively learn the importance of each channel and spatial position, thereby improving the model's classification accuracy. In the SAM spatial attention module, the convolution block in the original SAM channel was adjusted to use three 3×3 convolution kernels to generate spatial attention feature maps, which can more accurately identify the key information in the image and strengthen support for the classification task. Additionally, this study used transfer learning technology, which can use a pre-trained model for fine-tuning in a new task, thereby significantly reducing the model's training time and computational cost. This study selected weights from the ImageNet classification task for transfer, as ImageNet is a massive image dataset containing various categories of images, which can effectively improve the model's generalization ability. The main contributions of this article are as follows:

i) This study proposed a new attention mechanism module called CBAMp (convolutional block attention module pro) by improving the SAM spatial attention module in the CBAM attention mechanism. This innovation can improve the model's classification accuracy, especially for the classification of different lychee varieties with similar appearances, where it performs better.

ii) This study improved the ResNet-34 network structure using advanced techniques and further improved the model's performance by introducing the improved CBAMp attention mechanism, proposing an improved ResNet-34 residual network. This innovation can adaptively select the most discriminative and important regions in the image, thereby reducing the model's computational load and the impact of feature noise in redundant regions, thereby improving the accuracy and robustness of the model in lychee variety classification. It can solve the problem of difficult classification due to similar appearances of lychee varieties in the classification of lychee varieties.

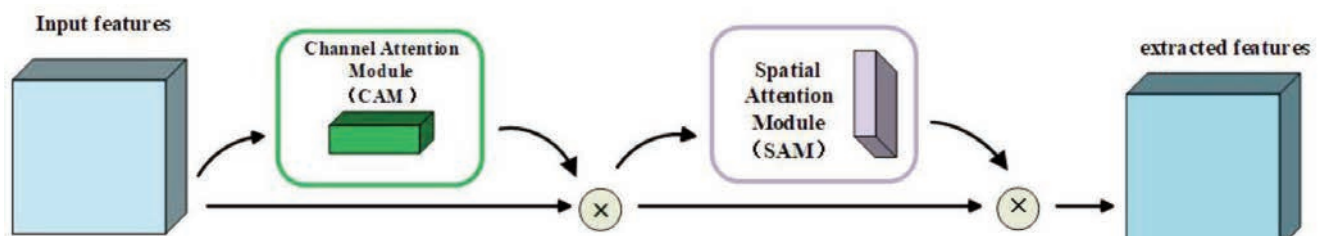iii) A series of simulation experiments are taken to prove the valid-

Figure 1. Overview of convolution block attention module.

ity of our suggested classification algorithm in this article compared with some general classification approaches.

Introduction to related knowledge
CBAM attention mechanism module

CBAM is a lightweight attention module proposed by Woo that can be embedded in convolutional neural networks (Woo *et al.*, 2018). Figure 1 show that the CBAM module consists of two independent sub-modules: the channel attention module (CAM) and the spatial attention module (SAM). CAM uses global average pooling to obtain global features for each channel. These global features are then mapped into a weight vector through two fully connected layers, which is multiplied with the input feature to obtain a weighted feature vector. SAM extracts spatial information through convolutional layers to generate a spatial weight matrix, which is then multiplied with the input feature to obtain a weighted feature vector.

Channel attention module

CAM is a channel attention mechanism that belongs to the hybrid attention mechanism, and its internal structure is shown in Figure 2. For an input feature map F, which is a four-dimensional tensor with shape (B, C, H, W), where B represents batch size, C represents the number of channels, and H and W represent the height and width of the feature map, respectively. After global average pooling (GAP) and global maximum pooling (GMP), it is compressed into two tensors with shapes of (B, C, 1, 1). These two operations capture the average response and maximum response within each channel, respectively, which helps to mine information between channels. Next, the two tensors obtained by GAP and GMP are respectively input into a multi-layer perceptron (MLP) with shared weights. This MLP usually consists of two fully connected layers with ReLU activation functions in between. As a result, two tensors are respectively transformed into two outputs with shapes of (B, C, 1, 1) through MLP. The outputs of GAP and GMP are element-wise added and then compressed to a value between 0 and 1 using the sigmoid activation function. The resulting channel attention weight tensor has a shape of (B, C, 1, 1). The
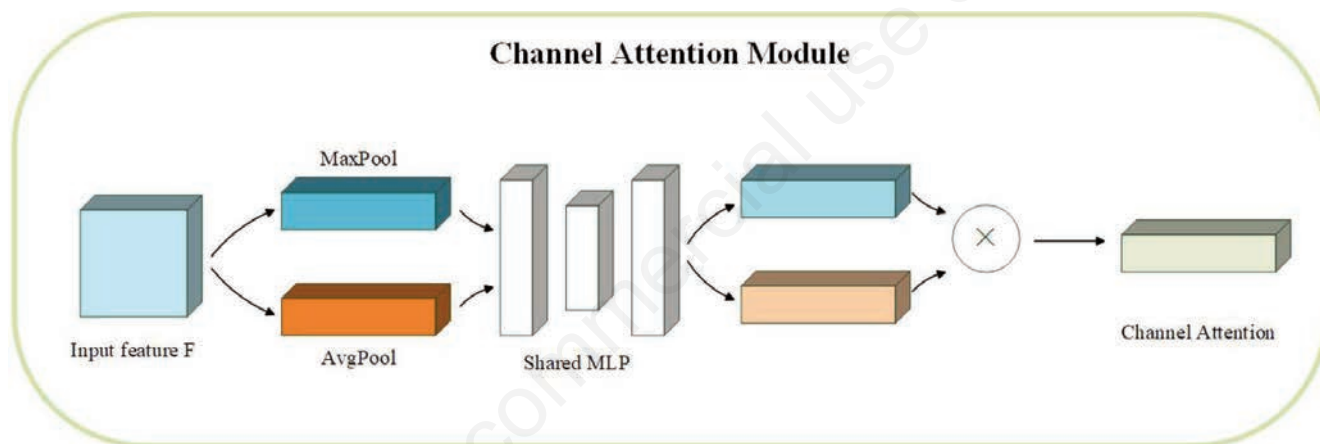

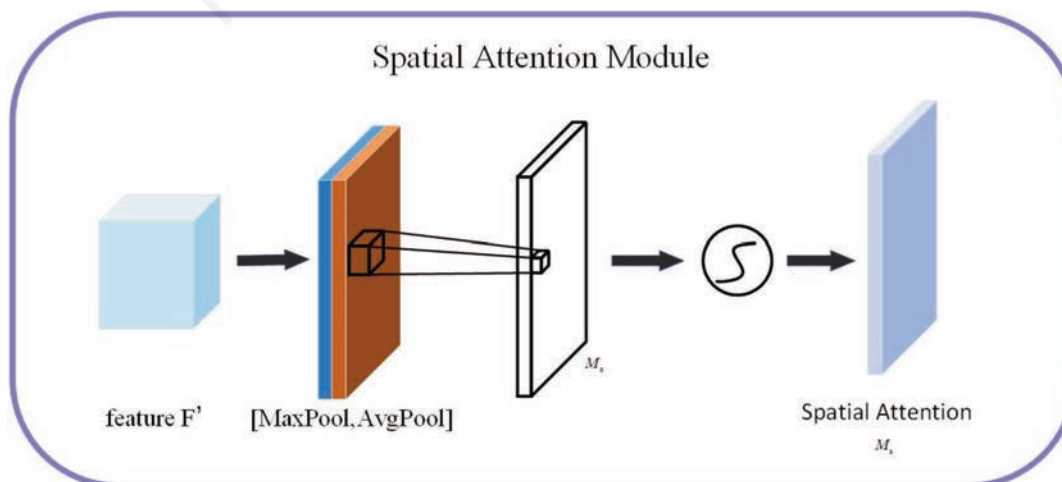**Figure 2.** Channel attention module.

Figure 3. Spatial attention module.

channel attention weight tensor is multiplied element-wise with the original input feature map to obtain an output feature map with a shape of (B, C, H, W). This weighting operation allows the network to automatically focus on channels that contain more representative information. Through this processing, CAM can enhance the channels in the input feature map that contain more representative information, thus improving the performance of the convolutional neural network.

Spatial attention module

SAM is a spatial attention mechanism belonging to the hybrid attention mechanism, and its internal architecture is shown in Figure 3. This module takes the feature map outputted by the CAM module as input and generates a spatial attention feature map for the spatial relationship inside the feature map. The purpose of the SAM network is to determine where the meaningful parts are in the feature map. The SAM network aggregates information from the input features using global maximum pooling and global average pooling, then concatenates the results of the two pooling operations and performs convolution to generate the spatial attention feature map. Finally, the spatial weight parameters are obtained through the sigmoid activation function. The calculation formula for SAM is shown in equation (1).

$$M_S(F) = \sigma(f^{7\times7}(AvgPool(F); MaxPool(F))) \tag{1}$$

In the above formula, F represents the input feature, AvgPool represents average pooling, MaxPool represents global pooling, σ is the Sigmoid activation function, and $f^{7\times7}$ represents a 7×7 convolutional layer.

ResNet-34 residual network

This study adopts ResNet-34 as the backbone network, which contains 34 convolutional layers and fully connected layers. The basic unit of ResNet-34 is the residual unit, which consists of two convolutional layers and a residual connection between them. By stacking different numbers of residual units, ResNet constructs networks of different depths, which can solve the problem of gradient vanishing in deep networks and adapt to different task

requirements (Lin *et al.*, 2023; Xuanjie *et al.*, 2023; Yu *et al.*, 2023). Figure 4 shows the ResNet-34 network model, with input images of size 224×224. The lychee images are fed into the ResNet-34 network, and the bottom-level features are extracted and dimension-reduced by a 7×7 convolutional kernel. These bottom-level features are then input to four sets of residual units to extract high-level features. The four sets of residual blocks can also increase the width and nonlinearity of the network.

Transfer learning

Transfer learning is a machine learning technique that allows us to apply knowledge learned on one task to another related task, thereby reducing the amount of training data and computing resources required (Shamshiri *et al.*, 2023; Taghizadeh and Hossein, 2023; Xuanyu *et al.*, 2023). Transfer learning has achieved significant success in the field of deep learning, especially in computer vision and natural language processing (Guarino *et al.*, 2023).

The basic idea of transfer learning is to use the knowledge learned from the source task to accelerate the learning of the target task. Typically, the source task and the target task have some degree of similarity, such as involving similar types of input data or having related objectives (Ni *et al.*, 2023).

The main steps of transfer learning in the field of deep learning are as follows: First, pre-training a deep learning model on the source task, during which the model learns many common features of the source task, which are largely related to the target task. This pre-trained model is usually called the base model. Next is fine-tuning, which involves fine-tuning the base model to adapt to the new task in the target task. For the case of limited target task data, most layers of the base model (usually convolutional or self-attention layers) are kept unchanged, the top classification layer is removed, and one or more new fully connected layers are added to adapt to the target task. In this case, only the newly added fully connected layers are trained, while the other layers of the base model remain unchanged. For the case of abundant target task data, most layers of the base model (usually convolutional or self-attention layers) are kept unchanged, the top classification layer is removed, and one or more new fully connected layers are added to adapt to the target task. In this case, we not only train the newly added fully connected layers, but also fine-tune part or all of the layers of the base model (Pan and Yang,, 2010).
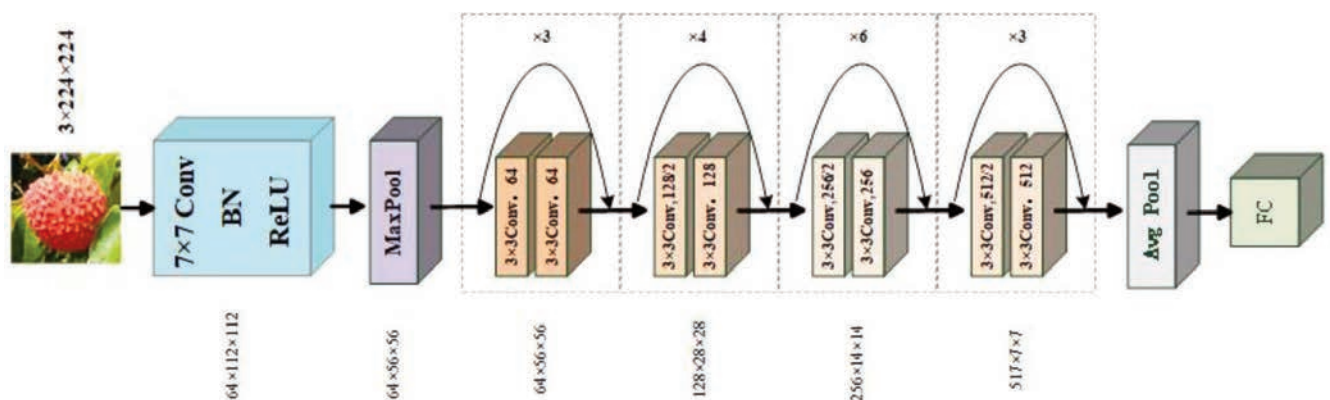

Figure 4. ResNet-34 backbone network.

Generally speaking, transfer learning techniques can bring less training data, faster training time, and better model performance.

Improvement method
Improved CBAM attention mechanism module

The SAM in CBAM is a method for extracting spatial features from images. It uses a 7x7 convolutional kernel to convolve with the results of global max pooling and global average pooling, in order to obtain effective spatial features. Using a large convolutional kernel can achieve a larger receptive field, but may lead to overfitting. This study proposes a new strategy that uses three 3x3 convolutional kernels in series instead of the original 7x7 convolutional kernel. By increasing depth to obtain the same size receptive field as the large convolutional kernel, it effectively prevents overfitting of the network. The innovation of this improvement method is to replace the original large convolutional kernel with a small kernel sequence to obtain the same size receptive field, reduce the number of network parameters, improve computational efficiency, and reduce the risk of overfitting. In addition, this improvement method can be combined with other convolutional neural network structures to improve the overall performance of the network. Specifically, it calculates the weight coefficients of the feature map through two convolutional operations and a sigmoid activation function, in order to obtain better image representations. This improvement method can play an important role in image recognition and other computer vision tasks, and has good prospects for promotion and application. The improved spatial attention calcula-tion formula is shown in equation (2).

$$M_S(F) = \sigma(f_{3\times3}^3(AvgPool(F); MaxPool(F))) \tag{2}$$

In the above equation, $f^3_{3\times3}$ represents three layers of 3x3 convolutions.

In this study, the original CAM module in CBAM was retained, and an improvement strategy was proposed for the SAM module. The improved CBAM was named convolutional block attention module pro (CBAMp). The overall architecture of CBAMp is shown in Figure 5.

Improved ResNet-34 residual network

The dataset used in this study has the characteristics of complex image backgrounds, a small proportion of effective information and the similarity of features among lychee varieties. In order to enhance the feature extraction ability of the network model in the face of complex and diverse lychee images, and to be able to focus on the differences in features between each variety, this study proposes a ResNet-34 model integrated with the innovative hybrid attention mechanism module CBAMp. The model utilizes the ability of CBAMp to comprehensively consider channel and spatial information. By modeling the information entering both the channel and spatial modules at the same time, the improved ResNet-34 model has stronger perceptual abilities, an improved understanding of the input images, and effectively obtains the overall contextual information of the lychee images.
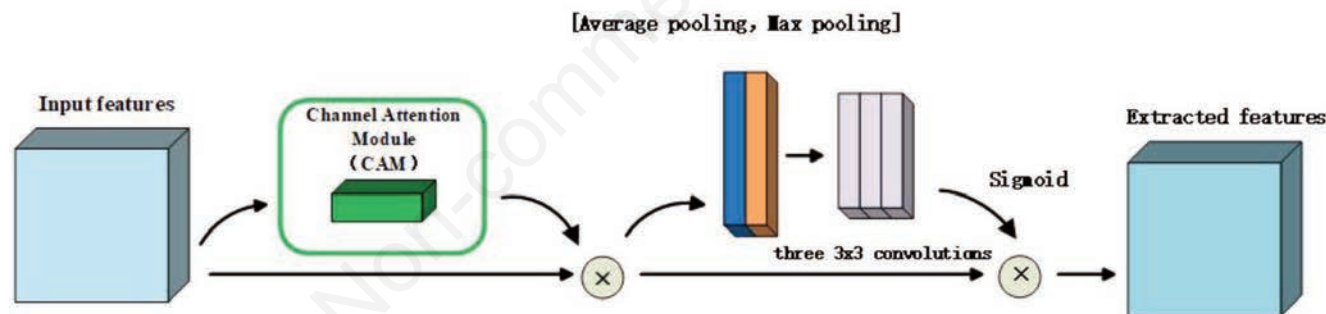
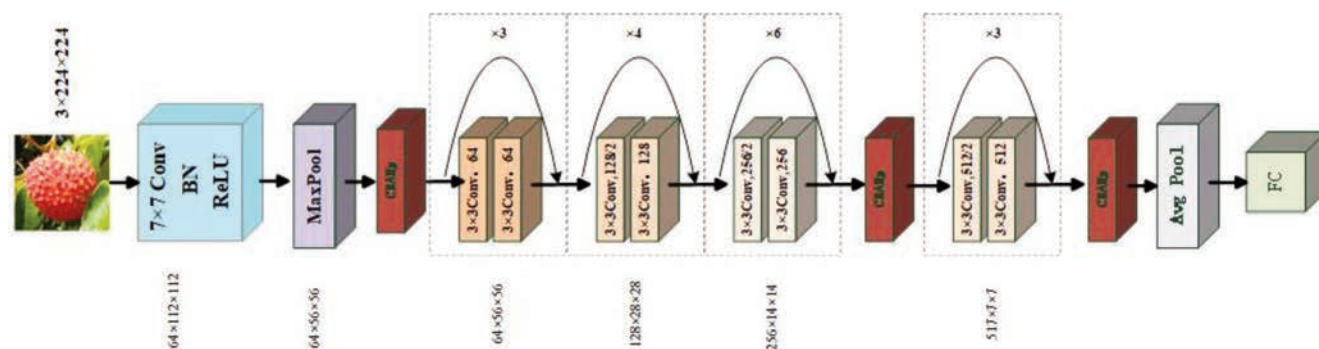

Figure 5. Overview of CBAMp.

Figure 6. Integrate CBAMp ResNet-34 model.

OPEN ACCESS

Figure 6 shows the overall network architecture of the ResNet-34 model with the hybrid attention mechanism CBAMp introduced in this study. The CBAMp module is added after the third stage following the max pooling layer and before the final global average pooling layer. The ResNet-34 network with CBAMp can better capture the spatial features of lychee fruit, such as shape, texture, and color, and focus on the correlation between information in different channels, thereby effectively extracting and utilizing the feature information of lychee fruit in the dataset.

Transfer learning usage

To improve the accuracy of lychee cultivar classification, this study adopted transfer learning. Specifically, the ResNet-34 model pre-trained on a large-scale image dataset was utilized, and its parameters were fine-tuned on the lychee dataset to adapt to its features. The process of transfer learning is shown in Figure 7. After transfer learning, the model's convergence speed was improved during training, and the risk of overfitting was reduced, thereby enhancing the accuracy of lychee cultivar classification. It is worth noting that in the process of transfer learning, it is necessary to choose a suitable pre-trained model and fine-tuning strategy based on different datasets and tasks, in order to fully utilize the features and knowledge learned from the pre-trained model, and improve the model's generalization ability and performance.

In this study, in order to apply transfer learning techniques, some adjustments were made to the task model:

Model structure adjustment. The top classification layer of the pre-trained model is removed and replaced with a new fully connected layer to adapt to the number of lychee varieties. This ensures that the output dimension of the model is consistent with the number of lychee varieties.

Fine-tune parameters. The pre-trained model is fine-tuned on the lychee dataset. In this study, most of the weights learned from the large-scale image dataset are preserved, but the weights of the top fully-connected layer and some other layers are updated to adapt to the lychee dataset's characteristics. This process can improve the model's performance on the target task by allowing it to leverage the generic feature representation learned from the source task while adapting to the specific requirements of the target task.

Experimental Results and Analysis

Experimental environment construction and parameter configuration

In the experiment, the computer system was Linux (Ubuntu20.04), with a 14-core Intel(R) Xeon(R) Gold 6330 CPU @ 2.00GHz and an NVIDIA RTX 3090 24GB graphics card. The deep learning framework used was PyTorch 1.10 with CUDA 11.3, and the memory was 64G.

The experimental settings in this study are as follows: the optimizer is set to SGD (Stochastic Gradient Descent), the number of epochs is 100, and the batch size for each input sample is 32, The images' input size is 224×224.

Experimental dataset

The dataset collection process in this study was divided into three parts. First, lychee images were gathered using web crawlers from search engines like Baidu and Google. Second, images were downloaded from websites such as Visual China and the Chinese Plant Image Database. Lastly, lychee images were collected through on-site photography. A total of six lychee varieties were collected, including 'Feizi Xiao', 'Guiwei', 'Huaizhi', 'Litchi Wang', 'Nuomi Ci', and 'Xian Jin Feng'. To ensure the accuracy of variety classification, the images collected from the internet underwent strict manual verification. In total, 3,799 original images were collected and divided into training and test sets at a ratio of 4:1. Table 1 provides detailed information about each lychee variety in the dataset, including the Chinese name, total number of images, number of images in the training and test sets, and examples of images for each variety.

Data augmentation

In this study, a dynamic data augmentation strategy was adopted, which involves performing image transformations in real-time each time an image is read from the dataset. The advantage of this approach is that even the same image may appear in different forms during different training iterations, significantly enhancing the diversity of the dataset (Liu and Wu, 2018; Li *et al.*, 2020).
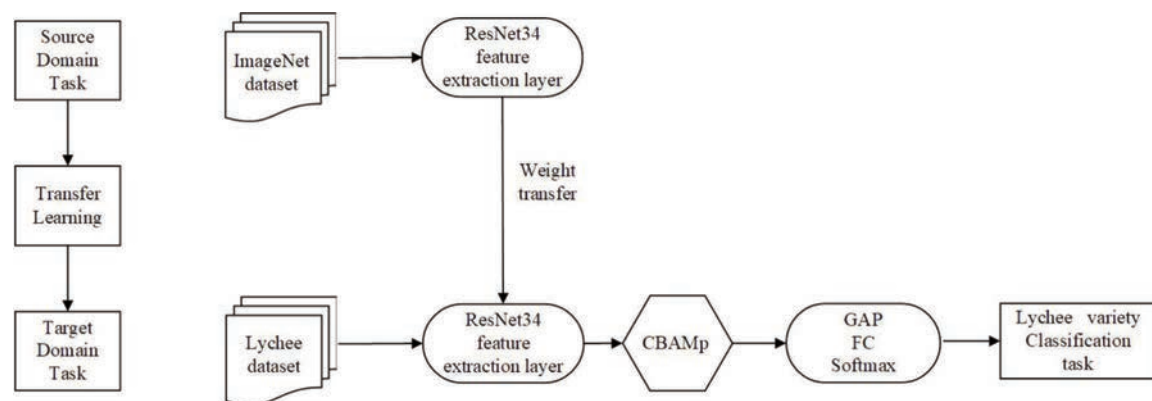

Figure 7. Resnet-34 network based on transfer learning and CBAMp fusion.

This diversification not only helps to prevent model overfitting but also enables the model to learn a broader range of feature representations, allowing it to adapt to new data outside of the training set.

The data augmentation operations applied in this study include mirror flipping, vertical flipping, adding Gaussian noise, random angle rotation, and center cropping. These operations are designed to simulate a variety of scenarios in actual applications, where the model may need to recognize images of different angles and qualities. The effects of the augmented images are shown in Figure 8.

For the training set, these augmentation operations are applied randomly each time a batch of data is loaded, aiming to increase the diversity of the dataset. Such diversity helps the model learn more generalized features, enabling it to better handle novel and previously unseen data. When processing the test dataset, random data augmentation operations are not used, as the test set is intended to represent the actual distribution of data, used for accurately assessing the model's performance.

Experimental methods and evaluation metrics

In this study, five classic convolutional neural networks were applied to the lychee classification task. The performance of the five networks was compared under the same optimizer (SGD), the same learning rate (lr=0.01), and the same number of iterations (epoch=100).

The VGG-16 and VGG-19 networks were designed by the Oxford Visual Geometry Group and use the same convolutional layer structure. The difference between the two networks lies in their depth and number of parameters. Both networks have good feature extraction abilities, but have a large number of parameters and are prone to overfitting, requiring more training time and computational resources. AlexNet was proposed by Alex Krizhevsky *et al.* in the ImageNet competition and was the first deep convolutional neural network model used in large-scale visual recognition tasks. AlexNet has multiple convolutional and pooling layers and

three fully connected layers were added at the end.

Recall, accuracy, precision and F1_score are used as evaluation metrics for the lychee variety identification model's classification accuracy on lychee RGB images, as shown in equations (3)-(6).

$$Recall = \frac{T_P}{T_P + F_N} \times 100\% \tag{3}$$

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \times 100\% \tag{4}$$

$$Precision = \frac{T_P}{T_P + F_p} \times 100\% \tag{5}$$

$$F1\_score = \frac{2 \times Precision \times Recall}{Precision + Recall} \times 100\% \tag{6}$$

In the equations, T_P represents the number of true positives (positive samples predicted as positive), F_P represents the number of false positives (negative samples predicted as positive), T_N represents the number of true negatives (negative samples predicted as negative), and F_N represents the number of false negatives (positive samples predicted as negative).

Experimental testing and analysis

Testing and analysis of classic convolutional neural networks

The five classic convolutional neural networks, namely VGG-16, VGG-19, AlexNet, ResNet-18, and ResNet-34, along with Vision Transformer and Swin Transformer, were applied to the task of classifying litchi varieties in this study. A comparison of their classification effectiveness was conducted. The comparison results are shown in Table 2.

From the experimental results in Table 2, it is evident that in the

Table 1. Basic information of lychee image data set.

Chinese name	Number of pictures/images	Number of training samples/images	Number of test samples/images	Example of experimental pictures
Feizixiao	600	480	120	
Guiwei	510	408	102	
Huaizhi	615	492	125	
LizhiWang	1004	801	203	
Nuomici	368	291	77	
Xianjinfeng	702	559	143	

task of classifying litchi varieties, VGG-19 scored the lowest in terms of accuracy and recall rates, at 72.4675% and 73.82% respectively. The ResNet networks significantly outperformed the others in both accuracy and recall rates. Among them, the best performing network was ResNet-34, achieving an accuracy and recall rate of 90.1344% and 91.23%, respectively, indicating its superior performance in the visual task of litchi classification. It is noteworthy that despite VGG-19 having a deeper model structure than VGG-16, its recognition accuracy decreased. The increased complexity and parameter count of the deeper model structure might lead to overfitting. For the VIT and Swin-T classification algorithms, which have performed well on the ImageNet dataset in the past two years, they achieved only 55.2981% and 47.5894% accuracy, respectively, in this study. This is attributed to the smaller scale of data in this research. Due to their larger number of parameters, higher complexity, and strong focus on global information, ViT and Swin-T might be more prone to overfitting on small datasets, leading to a decrease in accuracy compared to convolutional neural networks. Additionally, small datasets struggle to provide sufficient samples, making it difficult for complex models like Transformers to learn and generalize effectively, thus impacting their performance. In contrast, the ResNet network, utilizing residual block structures, allows the network to maintain high performance despite increased depth, and is less likely to suffer from gradient vanishing or exploding problems while learning higher-level features. However, due to the high similarity in texture, shape, and size features between some individual lychee varieties, the model's accuracy is only around 90%, which could result in many misclassifications with increasing data samples in practical applications. To address this, this study improved the ResNet-34 model to maximize the network's performance in lychee classification task.

Table 2. Comparison of classification algorithm performance.

Backbone network	Accuracy (%)	Recall (%)
VGG-16	78.7013	76.46
VGG-19	72.4675	73.82
Vison Transformer (VIT)	55.2981	51.74
Swin Transformer (Swin-T)	47.5894	49.31
AlexNet	80.2733	80.23
ResNet-18	89.2783	90.73
ResNet-34	90.2597	89.29

Table 3. The impact of various improvement strategies on the performance of the ResNet-34 network.

Model	Accuracy (%)	Recall (%)	F1_score (%)
ResNet-34	90.2597	89.2918	90.0410
Tran_ResNet-34	92.8571	92.7634	92.1274
ResNet-34_1C	88.9610	87.7438	88.0888
ResNet-34_1Cp	88.8312	87.4001	88.2937
Tran_ResNet-34_1C	94.5455	93.7460	94.4187
Tran_ResNet-34_1Cp	93.2468	91.8387	92.5686
Tran_ResNet-34_2C	95.3247	94.4453	94.9485
Tran_ResNet-34_2Cp	93.6364	94.1981	93.6747
ResNet-34_3C	91.5584	90.5332	91.2680
ResNet-34_3Cp	91.1688	89.9594	90.3629
Tran_ResNet-34_3C	95.7143	94.8778	95.3390
Tran_ResNet-34_3Cp	95.8442	95.2953	95.5561
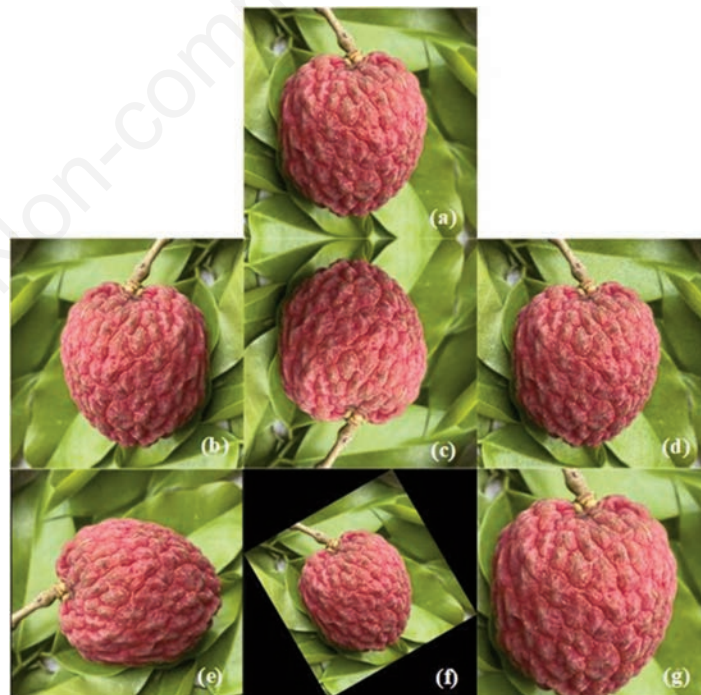
Figure 8. The effects of image enhancement after data augmentation. (**a**) is the original image, (**b**) - (**d**) are the effects of mirror flipping, vertical flipping, and adding Gaussian noise, respectively; (**e**) - (**f**) illustrate random angle flipping; and (**g**) is an example of center cropping.

Improved ResNet-34 testing and analysis

In this study, the ResNet-34 backbone network was improved by weight transfer, incorporating attention mechanism CBAM, and incorporating improved attention mechanism CBAMp. To verify the effectiveness of the improved network, the performance of the original ResNet-34 network was compared with the ResNet-34 network with various improvement measures applied, as shown in Table 3. In the table, the model naming convention is as follows: ResNet-34 represents the backbone network used; Tran represents the use of transfer learning technology; ResNet-34_1C with C represents the CBAM attention mechanism module, and 1C represents one CBAM attention mechanism module added after the Max Pool layer; ResNet-34_2Cp with Cp represents the improved attention mechanism module CBAMp, and 2Cp represents adding one CBAMp attention mechanism module after stage 3 on the basis of 1Cp; 3Cp represents adding a CBAMp module before Avg Pool layer on the basis of 2Cp.

From the results in Table 3, it can be seen that the proposed Tran_ResNet-34_3Cp performs the best among all the improvement measures, with an accuracy of 95.8442%, a recall of 95.2953%, and an F1_score of 95.5561%. In terms of accuracy, the improved network is 5.5845 percentage points higher than ResNet-34 in the lychee classification task. One of the main reasons is the introduction of transfer learning, through which the model can leverage knowledge from weights pre-trained on large datasets such as ImageNet. The key advantage of this technique lies in endowing the model with rich feature extraction capabilities during initialization, thereby reducing the amount of training data required. The performance improvement brought about by transfer learning stems from the model's enhanced generalization ability. Since the pre-training dataset covers a wide range of categories and image features, the model can learn richer feature representations. Subsequently, through fine-tuning, the model can better adapt to specific tasks, such as lychee classification. This approach not only reduces training time but also decreases the need for extensive data, enabling the model to perform better when faced with complex or diverse datasets. Additionally, the introduction of the CBAMp attention mechanism enables the model to more accurately focus on important feature information. Compared to traditional CBAM, CBAMp can more fully utilize feature information in lychee classification tasks, thereby improving classification accuracy and stability. To better understand and analyze the Tran_ResNet-34_3Cp model proposed in this paper, the Grad-CAM (gradient-weighted class activation mapping) model visualization technique was used to visually explain the effect of the improved model (Zhou *et al.*, 2016). This technique determines which part of the image contributes the most to the classification by calculating the importance of each class feature map to the output class. By inputting the trained weights, model information, and test images, a heatmap can be generated. Figure 9 shows several randomly selected lychee images from the dataset and the generated attention heatmap using Grad-CAM in this study. The more vibrant colors in the figure indicate the regions that the model pays more attention to.

The results from Figure 9 show that the proposed Tran_ResNet-34_3Cp model has a wider focus on the key areas of the lychee images compared to the original ResNet-34 model. It can recognize more features and more easily extract important characteristics that represent the lychee variety.

The confusion matrix between our proposed model and ResNet-34 is shown in Figure 10. The row labels of the confusion matrix represent the true labels of the predicted images, while the column labels represent the predicted labels of the model for the images. The values in the matrix represent the number of corresponding predicted lychee fruit images.

From Figure 10, it can be seen that compared to ResNet-34, the Tran_ResNet-34_3Cp model proposed in this paper has a more concentrated classification result on the diagonal of the confusion matrix, indicating that the model's lychee classification level is more accurate. By observing the incorrectly predicted samples, it can be seen that even with the improved network, 12 Xianjinfeng lychees were predicted as Nuomici, and 9 Litchi Kings were predicted as Xianjinfeng. This may be because Xianjinfeng is an improved variety of Nuomici, and some of its characteristics are inherited from Nuomici, resulting in similar appearance between the two varieties. Moreover, Litchi King and Xianjinfeng have similar contour shapes. Overall, the Tran_ResNet-34_3Cp model proposed in this paper has performed well in the lychee variety classification problem, and can effectively distinguish lychees and accurately identify their varieties.

Table 4 presents the recognition results of Tran_ResNet-34_3Cp for various lychee varieties. From the perspective of recall, F1 score, and precision, the model achieves or approaches 95% in most categories. This demonstrates the model's sensitivity to subtle features, allowing for accurate differentiation between different categories, indicating a high level of lychee recognition capability. However, the identification performance for lychee varieties such as Lychee King and Xianjinfeng is relatively poor. The main reason lies in the similarity of their contours, both exhibiting a heart-shaped appearance, and their skin texture resembling crocodile skin.

Impact of different attention mechanisms on lychee classification

To investigate the impact of different attention mechanisms on model performance, five sets of experiments were conducted, each incorporating SE, CA, ECA, CBAM, and the attention mechanism proposed in this paper, CBAMp. These mechanisms were added in the same positions as those in Tran_ResNet-34_3Cp. The results are presented in Table 5.

From the table, it is evident that the model's classification performance is optimal when incorporating the CBAMp attention mechanism proposed in this paper, achieving an accuracy of 95.84%. In contrast, the accuracies of SE, CA, ECA, and CBAM are relatively lower. From this experiment, it can be concluded that the CBAMp attention mechanism proposed in this paper is effective for lychee recognition.

Comparison of this study's method with existing advanced algorithms in the field of fine-grained image classification

This study's comparison results with other fine-grained image

Table 4. Classification performance of Tran_ResNet-34_3Cp on various lychee varieties.

Varieties	Recall (%)	F1_score (%)	Precision (%)
Feizi Xiao	100.00	99.59	99.17
Guiwei	99.02	97.12	95.28
Huaizhi	98.40	98.40	98.40
Nuomici	96.55	95.15	93.78
Lychee King	85.71	91.67	98.51
Xianjin Feng	90.21	90.53	90.85
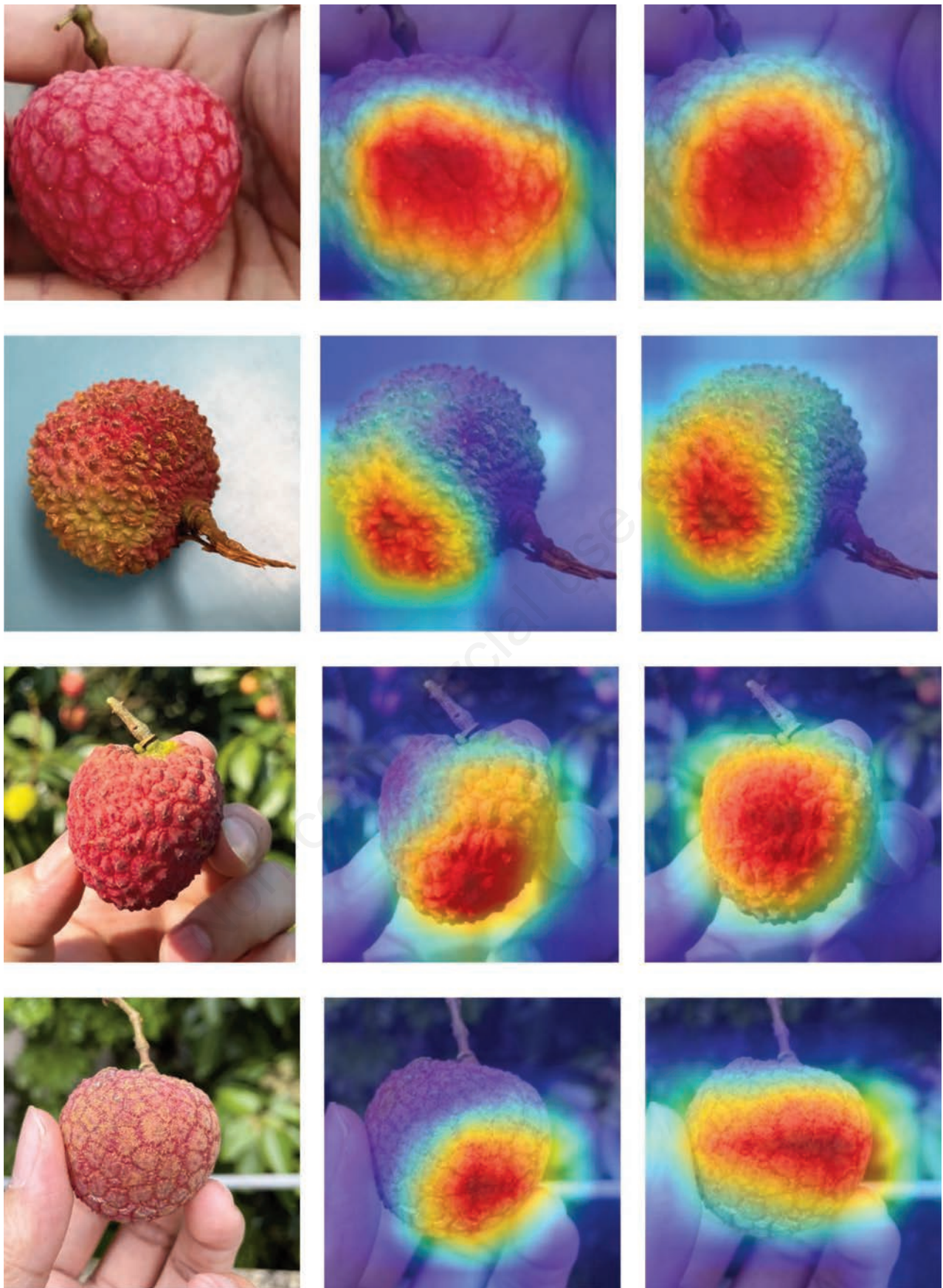
Figure 9. Heatmap for the model in this paper and ResNet-34.
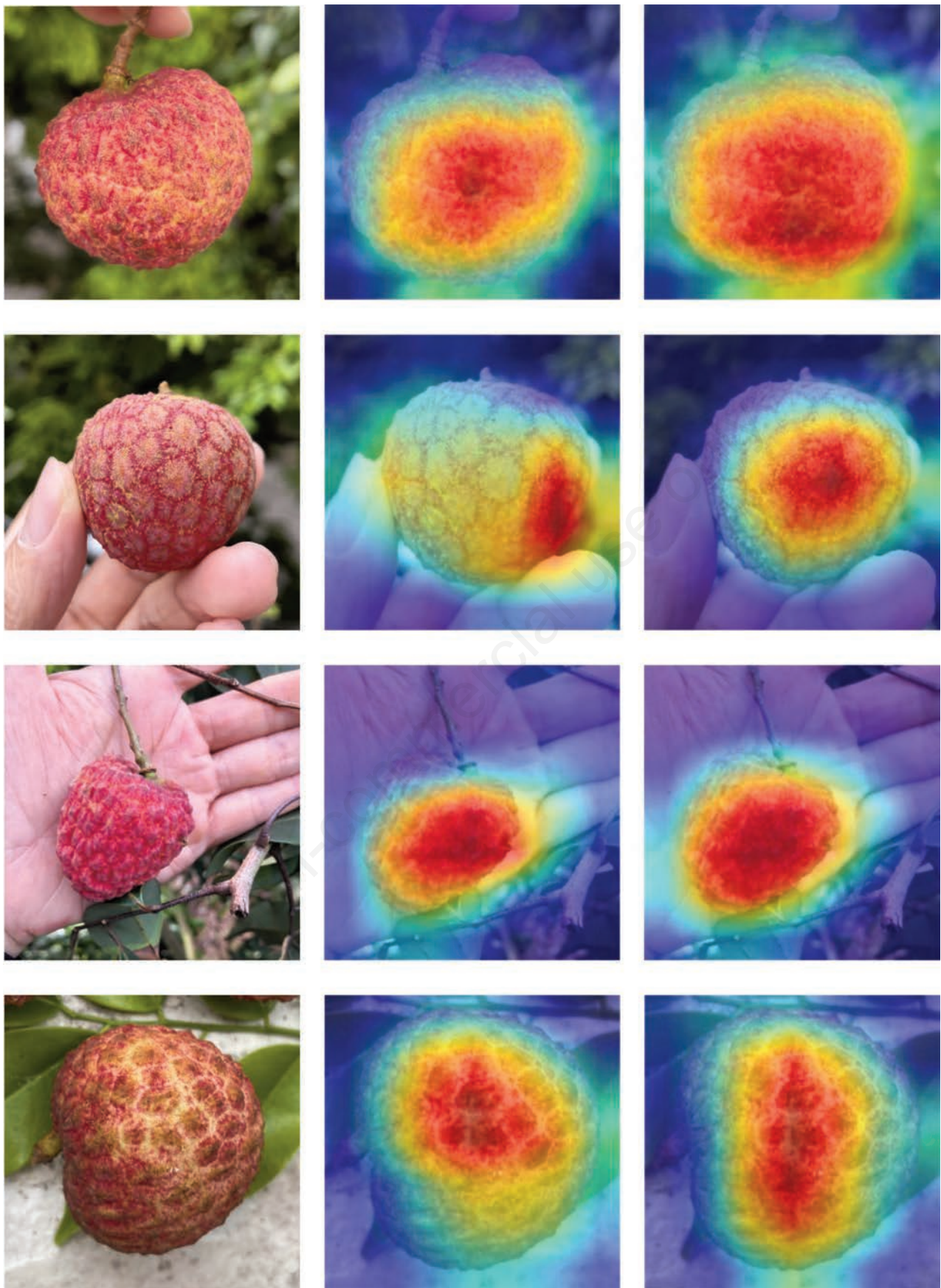
Figure 9. Heatmap for the model in this paper and ResNet-34.

classification algorithms are shown in Table 6. Stephen *et al.* (2023) and Wang *et al.* (2022) respectively studied the fine-grained classification of leaf diseases and insects, introducing self-attention mechanisms and residual attention blocks in their networks, making the models more suited for recognizing small object features. Sennan *et al.* (2022) constructed a spinach image dataset under laboratory conditions, while Wu *et al.* (2023) built a chicken image dataset in a farm environment. In contrast, the litchi image dataset used in this study was captured in complex natural settings.

In the field of litchi variety recognition, Osako *et al.* (2020) utilized the VGG-16 network to classify four types of litchi vari-

eties, achieving a high accuracy of 98.33%. However, the choice of litchi varieties in their study was limited and the image background was uniform, leaving the model's adaptability to real-world scenarios with complex backgrounds to be verified. This study applied the model published by Wu *et al.* (2023) for litchi image classification experiments, achieving only an 88.62% accuracy. This model was initially designed for identifying chicken gender and shows significant domain differences in recognizing specific visual features of litchi fruits, such as skin color and fruit shape. MPViT, one of the top-performing image classification algorithms at CVPR 2022, achieved only a 73.11% accuracy when applied
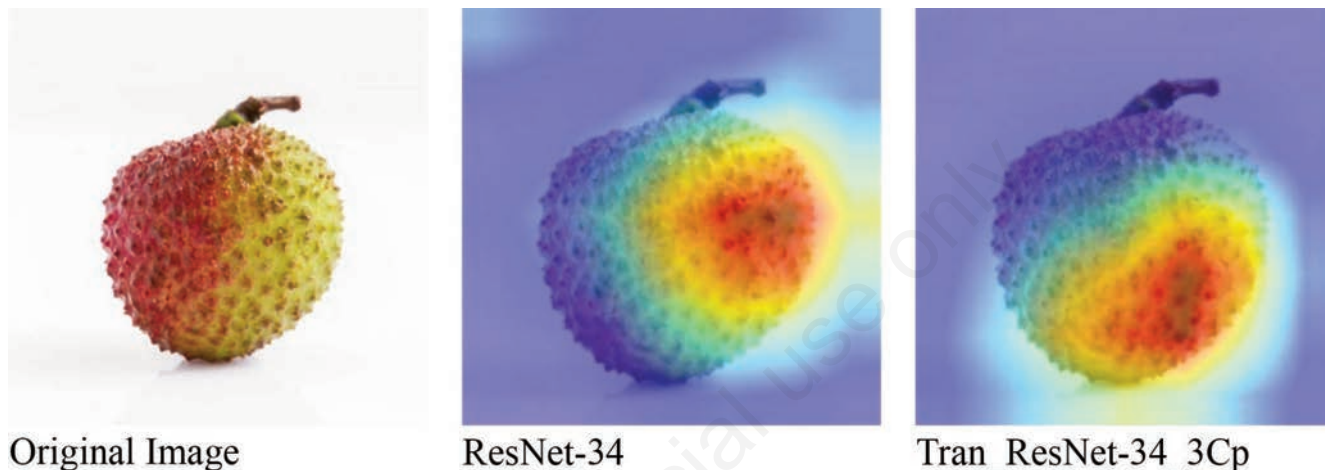
Figure 9. Heatmap for the model in this paper and ResNet-34.
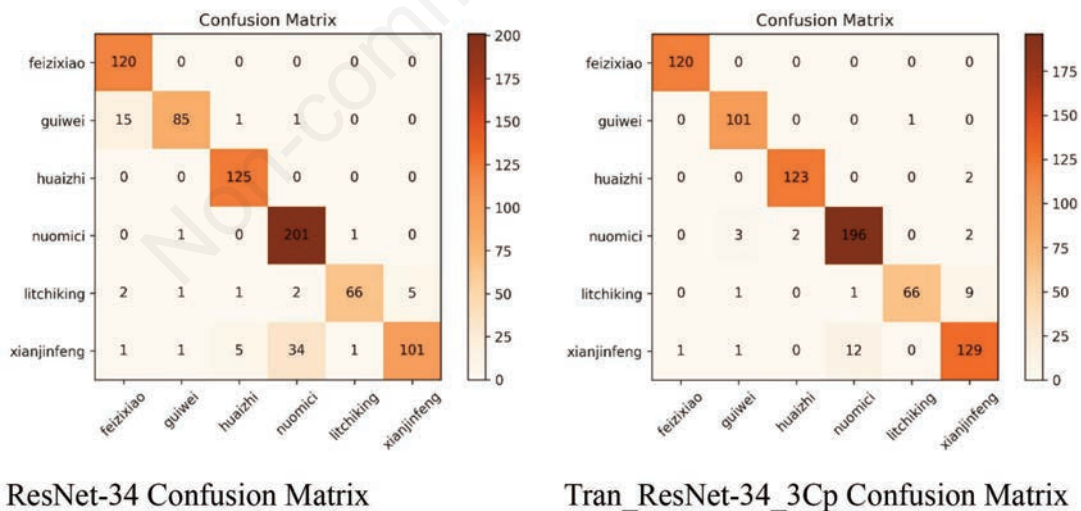
Figure 10. The Confusion Matrix of ResNet-34 andTran_ResNet-34_3Cp.

Table 5. The classification performance of different attention mechanisms..

Module	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5
ResNet-34	√	√	√	√	√
Attention mechanism	SE	CA	ECA	CBAM	CBAMp
Accuracy (%)	93.69	95.11	95.21	95.71	95.84

Table 6. A comparison of this research method with existing state-of-the-art algorithms in the field of fine-grained image classification.

| Paper	Object	Model	Types	Accuracy
Osako *et al.* (2020)	Litchi	VGG-16	4	98.33%
Stephen *et al.* (2023)	Rice leaf disease	ResNet-34+ self attention	4	98.54%
Wang *et al.* (2022)	Insects	S-ResNet	10	97.80%
Sennan *et al.* (2022)	Spinach	ResNet-50	4	98.70%
Wu *et al.* (2023)	Chicken	ResNet-50+SE+Swish	2	98.42%
Lee *et al.* (2022)	Litchi	MPViT	6	73.11%
Wu *et al.* (2023)	Litchi	ResNet-50+SE+Swish	6	88.62%
Our model	Lichi	Tran_ResNet-34_3Cp	6	95.84%

directly to this study's dataset, primarily due to the small scale of the dataset, which could not provide sufficient learning samples for MPViT (Lee *et al.*, 2022). This indicates that relying solely on the model's performance for assessment is insufficient; the compatibility of the algorithm with the application scenario also needs to be considered. Therefore, considering the practical limitations of data volume, the customized algorithm proposed in this study is the optimal choice for the current litchi variety recognition problem. In summary, the Tran_ResNet-34_3Cp model proposed in this paper performs best in the task of litchi classification, with its accuracy, recall rate, and F1_score all significantly surpassing the traditional ResNet-34 model, as well as outperforming the latest algorithms. The improvement in model performance is attributed to substantial modifications made to the traditional ResNet-34 in this study, enabling the model to better utilize feature information and focus on important features, thus enhancing feature representation and the model's generalization ability. This research provides an effective improvement strategy for the task of litchi classification and offers valuable insights and inspiration for the enhancement and optimization of other image classification tasks.

Conclusions

This study proposes a lychee variety identification method based on an improved ResNet-34, aiming to address the difficulties and subjectivity in lychee variety recognition in agriculture. The network structure of the existing ResNet-34 is improved by introducing the improved attention mechanism CBAMp, which enhances the network's ability to focus on the key features of different lychee varieties. On this basis, transfer learning technology is introduced. First, CBAM attention mechanism is introduced into the ResNet-34 network by adding CBAM after Max Pool, Stage3, and before Avg Pool, which enhances the network's ability to distinguish features between different categories while maintaining the original network architecture. Second, transfer learning is used to transfer the weights of a model trained on a large dataset to this task, improving model accuracy. Finally, the CAM channel in CBAM is retainxed, and the convolution kernel of the SAM channel is improved to deepen the network while maintaining the receptive field. Combining these improvements, the proposed model achieves high accuracy, effectively improving the model's recognition ability for lychee.

The method proposed in this paper demonstrates significant advantages in lychee recognition. While traditional identification methods at the molecular and genetic levels are accurate, they are complex to operate, cause sample damage, and are time-consum-

ing and costly. Although hyperspectral imaging combined with multivariate classification is non-destructive, it requires expensive equipment and is susceptible to environmental factors. In contrast, the deep learning method proposed in this paper does not require expensive instruments, resulting in lower research costs and no sample damage. By training deep learning models, this method can achieve rapid, effective, and economical lychee recognition while ensuring accuracy. It provides a solution to the complexity and potential damage risks associated with traditional methods.

The lychee varieties involved in this study only cover 6 common ones in Guangzhou market. According to incomplete statistics, there are as many as 200 lychee varieties, making it difficult to cover most of them. Therefore, the lychee varieties to be included in future research need to be expanded.

References

Akkem, Y., Biswas, S.K., Varanasi, A., 2023. Smart farming using artificial intelligence: a review. Eng. Appl. Artif. Intell. 120:105899.

Aradhya, M.K., Zee, F.T., Manshardt, R.M., 1995. Isozyme variation in lychee (litchi chinensis sonn.). Sci. Hortic. 63:21-35.

Chang, C.C., 1961. The lychee growing in taiwan. J. Agric. Assoc. China 33:51-63.

Chang, J., Chen, P., Chen, I., 2017. Litchi breeding and plant management in taiwan. The Lychee Biotechnology, 31-58.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I., 2021. MPViT: Multi-Scale Pyramid Vision Transformer for Dense Prediction Tasks. arXiv: 2112.11150.

Guarino, A., Malandrino, D., Zaccagnino, R., Capo, C., Lettieri, N. 2023. Touchscreen gestures as images. A transfer learning approach for soft biometric traits recognition. Expert Syst. Appl. 219:119614.

Jiang, H., Diao, Z., Shi, T., Zhou, Y., Wang, F., Hu, W., Zhu, X., Luo, S., Tong, G., Yao, Y., 2023. A review of deep learning-based multiple-lesion recognition from medical images: classification, detection and segmentation. Comput. Biol. Med. 157:106726.

Jiang, N., Zhu, H., Liu, W., Fan, C., Jin, F., Xiang, X. 2021. Metabolite differences of polyphenols in different litchi cultivars (Litchi chinensis Sonn.) Based on extensive targeted metabonomics. Molecules 26:1181.

Khurshid, S., Ahmad, I., Anjum, M.A., 2004. Genetic diversity in different morphological characteristics of litchi (Litchi chinensis Sonn.). Int. J. Agr. Biol. 6:1062-1065.

Lee, Y., Kim, J., Willette, J., Hwang, S.J., 2022. MPViT: multi-

path vision transformer for dense prediction. Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7287-7296.

Lin, K., Zhao, Y., Wang, L., Shi, W., Cui, F., Zhou, T., 2023. Mswnet: a visual deep machine learning method adopting transfer learning based upon resnet 50 for municipal solid waste sorting. Front. Env. Sci. Eng. 17:776.

Li, X., Huang, H., Zhao, H., Wang, Y., Hu, M., 2020. Learning a convolutional neural network for propagation-based stereo image segmentation. Vis. Comput. 36:39-52.

Liu, Y., Wu, Z., 2018. An improved threshold multi-level image recovery scheme. J. Inform. Secur. Appl. 40:166-172.

Liu, D., Wang, L., Sun, D., Zeng, X., Qu, J., Ma, J., 2014. Lychee variety discrimination by hyperspectral imaging coupled with multivariate classification. Food Anal. Meth. 7:1848-1857.

Liu, W., Xiao, Z., Bao, X., Yang, X., Fang, J., Xiang, X., 2015. Identifying litchi (litchi chinensis sonn.) Cultivars and their genetic relationships using single nucleotide polymorphism (snp) markers. Plos One 10:e01353908.

Madhou, M., Normand, F., Bahorun, T., Hormaza, J.I., 2013. Fingerprinting and analysis of genetic diversity of litchi (litchi chinensis sonn.) Accessions from different germplasm collections using microsatellite markers. Tree Genet. Genomes 9:387-396.

Menzel, C.M., Huang XuMing, H.X., Liu ChengMing, L.C., 2005. Cultivars and plant improvement. In: C. M. Menzel, G. K. Waite (eds.), Litchi and longan: botany, production and uses. Wallingford, CABI Publishing. pp. 59-86.

Mitra, S.K., Pathak, P.K., 2008. Litchi production in the Asia-pacific region. Acta Hortic. 863:29-36.

Ni, Z., Li, B., Yang, Y. 2023. Deep domain adaptation network for transfer learning of state of charge estimation among batteries. J. Energy Stor. 61:106812.

Osako, Y., Yamane, H., Lin, S., Chen, P., Tao, R., 2020. Cultivar discrimination of litchi fruit images using deep learning. Sci. Hortic. 269:109360.

Pan, S.J., Yang, Q. 2010. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22:1345-1359.

Sennan, S., Pandey, D., Alotaibi, Y., Alghamdi, S., 2022. A novel convolutional neural networks based spinach classification and recognition system. CMC-Comp. Mater. Cont. 73:343-361.

Shaikh, T.A., Rasool, T., Lone, F.R., 2022. Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming. Comput. Electron. Agric. 198:107119.

Shamshiri, M.A., Krzyżak, A., Kowal, M., Korbicz, J. 2023. Compatible-domain transfer learning for breast cancer classification with limited annotated data. Comput. Biol. Med. 154:106575.

Stephen, A., Punitha, A., Chandrasekar, A., 2023. Designing self attention-based resnet architecture for rice leaf disease classification. Neural Comput. Appl. 35:6737-6751.

Taghizadeh, A.A.A., Hossein, M., 2023. A novel application of deep transfer learning with audio pre-trained models in pump audio fault detection. Comput. Ind. 147:103872.

Woo, S., Park, J., Lee J-Y., Kweon, S., 2018. CBAM: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – ECCV 2018. Lecture Notes in Computer Science, vol 11211. Cham, Springer.

Wang, P., Luo, F., Wang, L., Li, C., Niu, Q., Li, H., 2022. S-resnet: an improved resnet neural model capable of the identification of small insects. Front. Plant Sci. 13:5241.

Wang, T., Zhao, L., Huang, P., Zhang, X., Xu, J., 2021. Haze concentration adaptive network for image dehazing. Neurocomputing 439:75-85.

Wu, D., Ying, Y., Zhou, M., Pan, J., Cui, D., 2023. Improved resnet-50 deep learning algorithm for identifying chicken gender. Comput. Electron. Agric. 205:107622.

Wu, S.X., 1998. Encyclopedia of China fruits: litchi. China Forestry Press, Beijing.

Xuanjie, Q., Fang, Y., Haihong, L., 2023. A difference attention ResNet-LSTM network for epileptic seizure detection using EEG signal. Biomed. Signal Process. Control 83:104652.

Xuanyu, W., Yixiong, F., Shanhe, L., Hao, Z., Bingtao, H., Zhaoxi, H., Jianrong, T., 2023. Improving NeuCube spiking neural network for EEG-based pattern recognition using transfer learning. Neurocomputing 52:222-235 9.

Yao, P., Gao, Y., Simal-Gandara, J., Farag, M.A., Chen, W., Yao, D., Delmas, D., Chen, Z., Liu, K., Hu, H., 2021. Litchi (litchi chinensis sonn.): A comprehensive review of phytochemistry, medicinal properties, and product development. Food Funct. 12:9527-9548.

Yu, H., Liu, J., Chen, C., Heidari, A.A., Zhang, Q., Chen, H., 2022a. Optimized deep residual network system for diagnosing tomato pests. Comput. Electron. Agric. 195:106805.

Yu, H., Sun, H., Tao, J., Qin, C., Xiao, D., Jin, Y., Liu, C., 2023. A multi-stage data augmentation and ad-resnet-based method for epb utilization factor prediction. Autom. Constr. 147:104734.

Yu, H., Yan, X., Liu, S., Li, P., Hao, X. 2022b. Radar emitter multi-label recognition based on residual network. Def. Technol. 18:410-417.

Zhang, R., Zeng, Q., Deng, Y., Zhang, M., Wei, Z., Zhang, Y., Tang, X., 2013. Phenolic profiles and antioxidant activity of litchi pulp of different cultivars cultivated in southern China. Food Chem. 136:1169-1176.

Zhao, Y., Wang, X., Che, T., Bao, G., Li, S., 2023. Multi-task deep learning for medical image computing and analysis: a review. Comput. Biol. Med. 153:106496.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A. 2016. Learning deep features for discriminative localization. Proc. IEEE Conf. on Computer Vision and Pattern Recognition. pp. 2921-2929.