Fine-grained classification of grape leaves via a pyramid residual convolution neural network

Hanghao Li[1], Yana Wei[1], Hongming Zhang[1*], Huan Chen[1], Jiangfei Meng[2]

(1. *College of Information Engineering, Northwest A&F University, Yangling 712100, Shaanxi, China*;
2. *College of Enology, Northwest A&F University, Yangling 712100, Shaanxi, China*)

Abstract: The value of grape cultivars varies. The use of a mixture of cultivars can negate the benefits of improved cultivars and hamper the protection of genetic resources and the identification of new hybrid cultivars. Classifying cultivars based on their leaves is therefore highly practical. Transplanted grape seedlings take years to bear fruit, but leaves mature in months. Foliar morphology differs among cultivars, so identifying cultivars based on leaves is feasible. Different cultivars, however, can be bred from the same parents, so the leaves of some cultivars can have similar morphologies. In this work, a pyramid residual convolution neural network was developed to classify images of eleven grape cultivars. The model extracts multi-scale feature maps of the leaf images through the convolution layer and enters them into three residual convolution neural networks. Features are fused by adding the value of the convolution kernel feature matrix to enhance the attention on the edge and center regions of the leaves and classify the images. The results indicated that the average accuracy of the model was 92.26% for the proposed leaf dataset. The proposed model is superior to previous models and provides a reliable method for the fine-grained classification and identification of plant cultivars.

Keywords: fine-grained classification, grape cultivars identification, pyramid residual network, convolution neural network
DOI: 10.25165/j.ijabe.20221502.6894

Citation: Li H H, Wei Y N, Zhang H M, Chen H, Meng J F. Fine-grained classification of grape leaves via a pyramid residual convolution neural network. Int J Agric & Biol Eng, 2022; 15(2): 197–203.

1 Introduction

Grapes are one of the most valuable horticultural crops in the world. The statistical database of the Food and Agriculture Organization of the United Nations (FAO), states that grape production was 77 million t in 2019. China produces one of the most grapes globally[1]. The value of different cultivars of table grapes varies greatly, and the cultivar also greatly affects the quality of the wine. Transplanted grape seedlings require several years to bear fruit, but leaves can mature in a few months. The early identification of cultivars could thus be used to judge whether a cultivar was suitable for local planting, which could reduce the waste of resources. Mixtures of different grape cultivars can negate the benefits of improved cultivars and hamper the protection of genetic resources and the identification of new hybrid cultivars. An automated system of identification would free experts from the task of routine identifications and allow them to focus on the more conceptually difficult issues of discovering, describing, and revising species concepts[2].

The classification and identification of leaves have recently received much attention. The methods used can be divided into two types, image analysis, and learning-based methods. Image analysis generally uses mathematical models and image processing technology to analyze features for extracting useful information from images. Yousefi et al.[3] proposed the Rotation Invariant Wavelet Descriptors, a new shape descriptor. They used a multilayer perceptron neural network as a classifier and evaluated their approach using Flavia[4], a publicly available standard dataset. Saleem et al.[5] used twenty handcrafted shape parameters extracted from leaf images for extracting features and used principal component analysis to remove redundancy. They proposed a five-step algorithm to recognize the plant type and evaluated the algorithm using Flavia and their dataset. Xue et al.[6] used the automated image analysis and visible/near-infrared spectral based parameters obtained from leaves as inputs to construct customized artificial neural network models classifying twenty kinds of medicinal plants. Wang et al.[7] proposed a novel multi-scale sliding chord matching method for identifying leaf images of soybean cultivars. They divided the leaves into three subsets from the lower, middle, and upper parts of the soybean plant and tested them with single and combined leaf patterns. These image analysis studies improved the classification accuracy by proposing a new leaf feature descriptor or a feature matching method. These methods are less dependent on the amount of dataset than learning-based methods. Some issues, however, remain, despite the success of the classical methods of image analysis. These algorithms rely on manually extracted leaf features, which are usually designed based on manual screening, so capturing high-level semantic features and complex content is difficult. Each step of the algorithm is also relatively independent and lacks global optimization schemes.

Traditional learning-based methods use algorithms to select features and then classify them using machine learning. The deep learning of automatic extracting features using neural networks,

Received date: 2021-07-16 **Accepted date:** 2022-03-03
Biographies: Hanghao Li, Master, research interests: computer vision, pattern recognition, Email: emmawatson@nwafu.edu.cn; **Yana Wei**, Master, research interests: image recognition, Email: viana1207@163.com; **Huan Chen**, PhD, research interests: evolutionary computation, remote sensing image change detection, Email: huanchen@nwafu.edu.cn; **Jiangfei Meng**, PhD, Associate Professor, research interests: viticulture and grape breeding, Email: mjfwine@nwafu.edu.cn.
***Corresponding author: Hongming Zhang**, PhD, Professor, research interests: artificial intelligence, intelligent agriculture, remote sensing. College of Information Engineering, Engineering, Northwest A&F University, Yangling 712100, Shaanxi, China. Tel: +86-29-87091197, Email: zhm@nwsuaf.edu.cn.

however, has more advantages due to its simplicity and efficiency[8,9]. In the study of leaf classification based on deep learning, Hall et al.[10] introduced a variety of conditions to the test dataset and combined hand-extracted features, a histogram of curvature scales, a deep CNN, and random forest classifiers to classify the Flavia dataset, achieving good accuracy. The Flavia dataset, however, was mainly aimed at different kinds of leaves with large differences in shape. Studying the fine-grained classification of leaf images of the same kind and different species is more challenging. Pereira et al.[11] evaluated techniques of transfer learning and fine-tuning techniques based on the AlexNet network and achieved 73% test accuracy in the classification of leaves of six grape cultivars. Yang et al.[12] proposed scanning leaf patches around the central blades and then inputting the leaf or leaf spot into the classifier based on the deep CNN model VGG16, Inception-V3, and NASNet to distinguish three kinds of plants with morphologically similar leaves. Kaya et al.[13] investigated the effects of four transfer learning models on deep neural network-based plant classification on four public leaf datasets. Tavakoli et al.[14] proposed a method for changing the classifier of the last layer of VGG16[15] to classify the upper and lower surface datasets containing three species of leaves. They trained models for classifying species (three classes), cultivars of the same species (four classes), and cultivars of different species (twelve classes) and compared their classifier with Softmax, NormFace[16], CosFace[17], and ArcFace[18]. Compared with image analysis methods, learning-based methods are simpler because features can be extracted automatically, and the models are better for classification. However, these methods require high computational power and require more images as training datasets. These studies have made some progress in the classification accuracy of different kinds of leaves, but the classification of cultivars of the same species based on leaf images can still be improved.

Multi-scaling uses multiple branches, where each branch deals with a feature at one scale. This method can obtain multi-scale information for the target and increase the receptive field of the network. Lin et al.[19] proposed a top-down architecture with horizontal connections for constructing high-level semantic feature maps at various scales. In this model, a feature extractor called a feature pyramid network was constructed using the inherent multi-scale and pyramid hierarchical structure of deep convolution networks, obtaining good single-mode results for the COCO detection benchmark. Meng et al.[20] proposed an adaptive resolution network called AR-NET combined with a strategy network to select the optimal resolution for each input image. The effectiveness of the method was demonstrated using extension experiments of the network on several action recognition benchmark datasets.

The goal of this study was to improve the accuracy of fine-grained classification of different cultivars of grape leaf classification for precise vineyard management and to provide a reference for the classification of other plant species. In this study, a pyramid residual convolution neural network that can extract multi-scale feature images of leaves based on morphological characteristics is introduced. The network improves the model's attention to the edge and center of the leaf by combining leaf multi-scale features. The proposed model and the commonly used convolution neural network classification models were evaluated using a dataset of leaves of 11 grape cultivars. In addition, a public agricultural image classification dataset PlantVillage[21] was used to check the generalization of the proposed approach.

2 Material and methods

2.1 Dataset of leaf images

A dataset of mature leaves from eleven grape cultivars was established. These leaves were collected on sunny skies in July, over two days. More than 50 leaves of each cultivar were collected from an experimental field at Northwest A&F University. The leaves were preserved intact and brought back to the laboratory. After rinsing, they were photographed using a fixed shooting table to ensure the relative size of the leaves in the images was consistent, at a resolution of 3000 pixels×3000 pixels for each image. A HUAWEI Mate 20 mobile phone was used to take the images, which were saved in JPEG format. Both the upper and lower sides of the leaves were captured. The dataset includes 1115 images in total and Figure 1 presents sample images from the dataset along with the number of images for each cultivar.

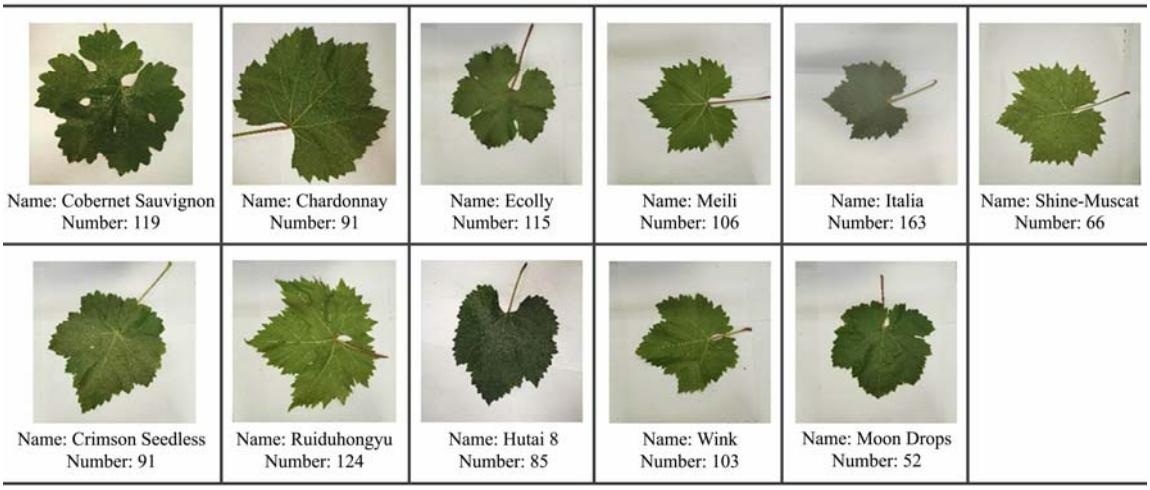


Figure 1 Mature grape leaves of eleven cultivars

2.2 Dataset preprocessing

The bilinear interpolation was used in this study to reduce the resolution of the original images to 224×224. 80% of all leaves were used as a training dataset, and the remaining 20% were used as a test dataset. 20% of the data in the training set were randomly used as the validation set to calculate the loss function during training, and the test dataset was only used to evaluate the performance of classifiers. The images in both the training and testing datasets were enhanced by rotating 90°, flipping left and right, flipping up and down, and randomly adjusting brightness,

contrast, hue, and saturation. The enhancements helped to reduce the risk of overfitting and improved the ability to generalize the classifier, thereby improving the robustness of the model. The training and testing datasets have 6244 images and 1561 images after data enhancement, respectively.

2.3 Pyramid residual convolution neural network

The shape and venation of the grape leaves played an important role in identification[22,23], so they were considered to be important regions. When using the convolution layer to extract leaf features, two feature maps were visualized and extracted using Grad-cam[24] and found that these regions of the model were concentrated in the center and at the edges of the leaves (Figure 2). Therefore, the pyramid structure was used to construct three parallel backbone classification networks and used the fused features to direct the attention of the model to these important regions.

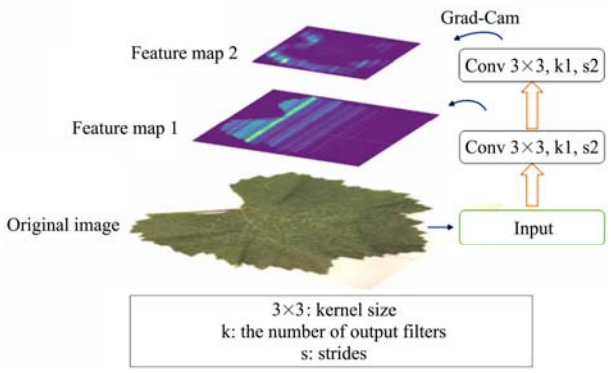
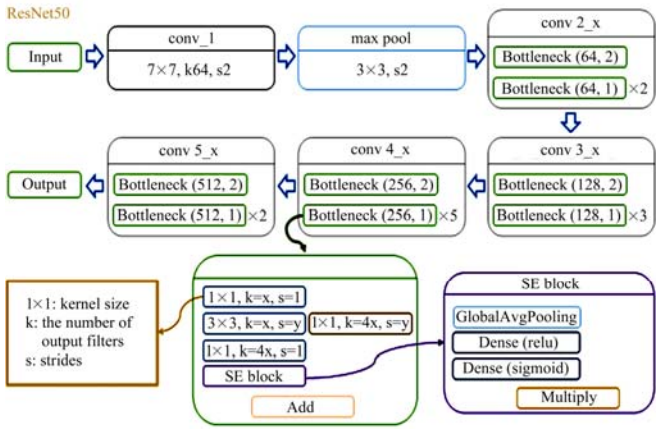


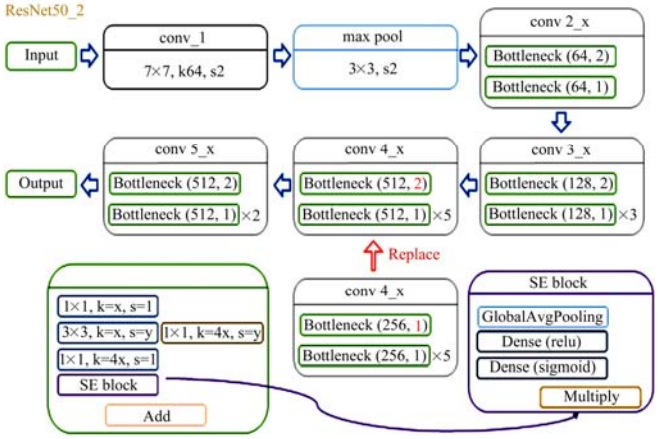
Figure 2 Two feature maps obtained by convolution

The network in this study was based on the ResNet50 CNN structure as the backbone network, which is widely used and performs well in the field of classification. The network mainly consists of 49 convolution layers and one maximum pooling layer. The squeeze-and-excitation (SE) attention mechanism[25] was added to the convolution module of ResNet to improve the spatial performance of the network. Its structure is shown in Figure 3a.

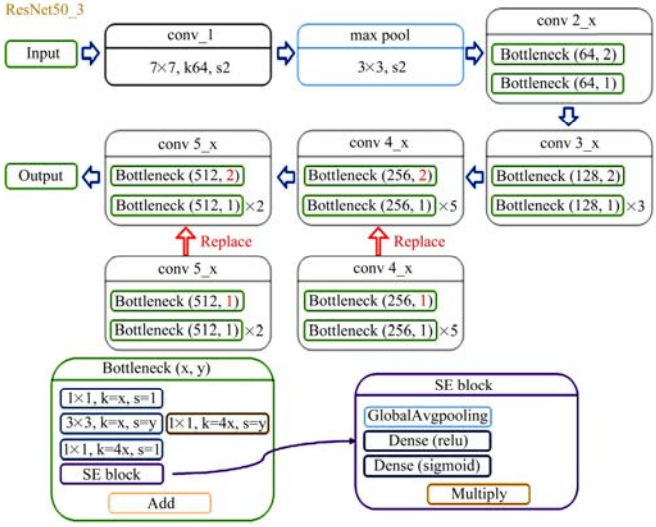
Two connected convolution layers with a stride size of 2 were used (Figure 2), to obtain two feature maps with original image sizes of 1/2 and 1/4. Semantic information from different scales was obtained using the pyramid residual CNN model. The model that used the original image and feature map 1 was called PyramidTwoResNet50se, and the model that used the original image and two feature maps were called PyramidTriResNet50se. The "se" at the end of these model names indicates that the models added the SE attention mechanism[25].



a. ResNet50 internal structure



b. ResNet50_2 internal structure



c. ResNet50_3 internal structure

Figure 3 ResNet50 backbone model and modified ResNet models

These feature maps and the original image were used as the inputs of three ResNet50 models. The stride of some convolution layers was changed to 1 of the residual networks, with the feature map as input, so the output shapes of the three networks were consistent. The modified ResNet model is displayed in Figures 3b and 3c. The classification model of the pyramid residual structure was then established using the feature-fusion method of adding the convolution kernel outputs of the three networks and adding Dropout and a dense layer with the number of classification nodes. Based on the backbone model, the pyramid models directed attention to the important regions. The pyramid residual CNN PyramidTriResNet50 model is shown in Figure 4.

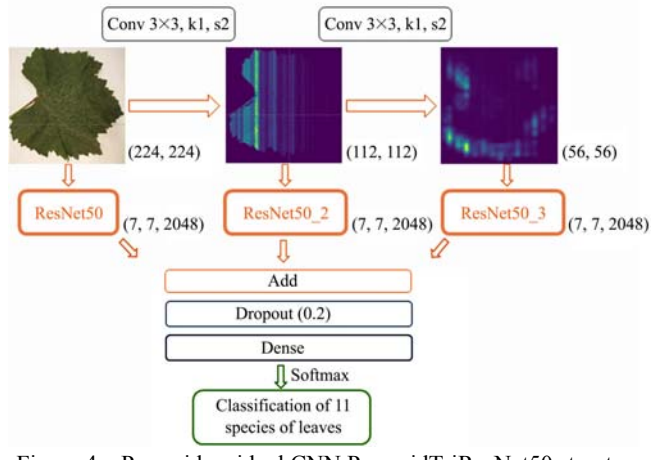


Figure 4 Pyramid residual CNN PyramidTriResNet50 structure

2.4 Training procedure and performance evaluation

All experiments were implemented using Tensorflow 2[26]. The models were trained using a single NVIDIA Tesla P40 GPU with 20 GB of memory. The hyperparameters were set with adam as an optimizer, a learning rate of 10^{-4}, a learning-rate decay of 10^{-7}, and a batch size of 200 and 300 training iterations. Several CNNs were implemented, including general classification CNNs and CNNs for fine-grained leaf classification, and tested them using our leaf data set. In the selection of comparative model parameters, the loss function classifier of the model proposed by et al.[14] refers to the optimal value in the original research. The optimizer sets the parameter of the combined loss function to the stochastic gradient descent algorithm, where the parameters of the combined loss function are m_1=0.35 and m_2=0.10, the parameter of the ArcFace loss function is m=0.5 and the parameter of the CosFace loss function is m=0.4. The 5-fold cross-validation was used to calculate the average accuracy of these models. Various metrics were calculated including Precision, Recall, and F1-score based on confusion matrices to evaluate the performance of the proposed model and the overall classification accuracy[14].

$$\text{Accuracy} = \frac{\text{Number of recognized samples}}{\text{Number of total samples}} \times 100\% \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \times 100\% \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \times 100\% \quad (3)$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\% \quad (4)$$

where, Accuracy is used to evaluate the quality of the model, which represents the proportion of samples with correct classification to the whole sample; Precision indicates how many positive samples are predicted to be positive according to the prediction results; Recall indicates how much of the original positive samples are predicted correctly; F1-score is the harmonic mean of Precision and Recall, combining these two indicators; TP, TN, FP, and FN are the numbers of true positives, true negatives, false positives, and false negatives, respectively. When calculating metrics of a class, the other classes in the dataset are considered negative.

Table 1 Classes selected in the PlantVillage dataset and the number of images corresponding to each category in training

Category name	Images selected	Enhancement	Images for training
Apple___healthy	683	×1	683
Blueberry___healthy	701	×1	701
Cherry_(including_sour)___healthy	684	×1	684
Corn_(maize)___healthy	694	×1	694
Grape___healthy	339	×2	678
Orange___Haunglongbing_ (Citrus_greening)	676	×1	676
Peach___healthy	288	×2	576
Pepper,_bell___healthy	711	×1	711
Potato___healthy	121	×5	605
Raspberry___healthy	297	×2	594
Soybean__healthy	706	×1	706
Squash___Powdery_mildew	700	×1	700
Strawberry___healthy	364	×2	728
Tomato___healthy	738	×1	738

The generalization ability of the proposed model was evaluated using an agricultural image classification dataset, PlantVillage[21]. Fourteen kinds of plant leaves of different species were selected for classification because this study aimed to classify different cultivars based on leaves, while the other categories were aimed at classifying different diseases of the same cultivar of leaves. The number of selected training dataset images is listed in Table 1. The dataset was balanced by enhancing the classes with fewer images than others, and then randomly divided into two parts in proportion to 3:1 for training and validating, respectively. All models were trained using the same hyperparameters and the model with the lowest loss of validation during training was used to classify the test dataset and evaluate the performance of the model. There is a portion of the PlantVillage dataset to test the accuracy of classification models and the test dataset for each of these fourteen categories is 200 images.

3 Results and discussion

3.1 Selection of backbone network

In the selection of the backbone network, generic and high-performing classification models were tested to find a suitable backbone model for classifying the grape leaves (Table 2). These models were specified with uniform hyperparameters, and training was stopped when the loss of validation was stable. The test results indicated that ResNet50 and ResNet101 performed best. ResNet50 was selected as the backbone network because its average accuracy was high and required fewer computing resources.

Table 2 Performance of image classification of the CNN models for the dataset of grape leaves

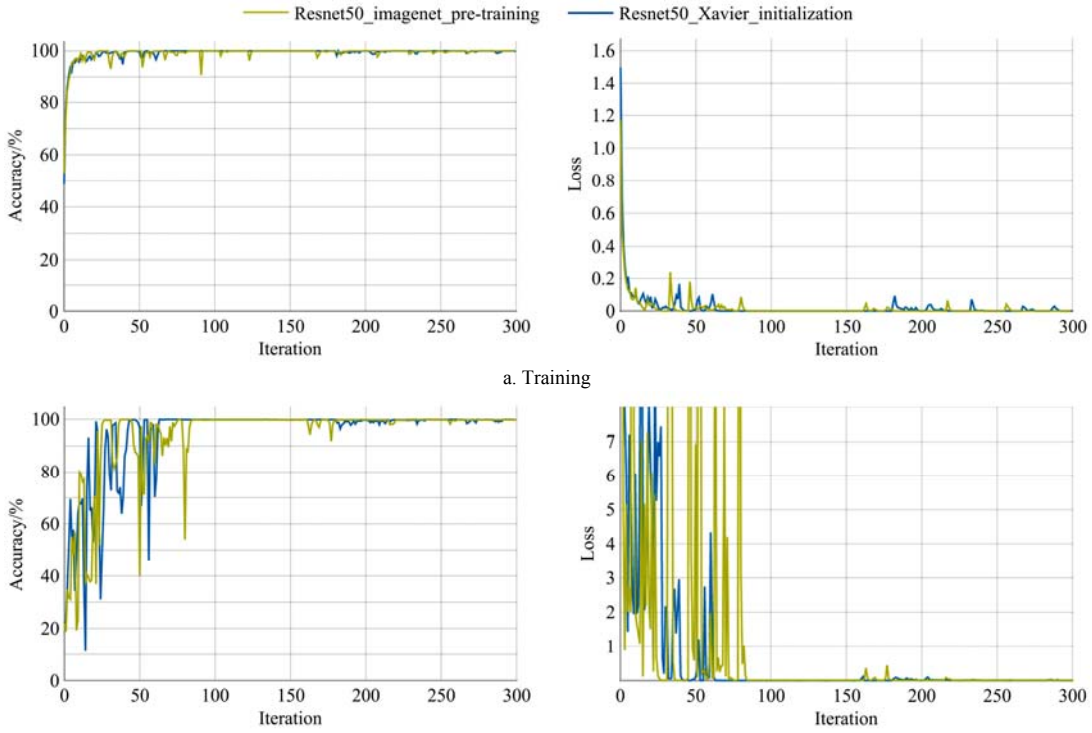
Model	Accuracy/%
VGG16	72.62±2.30
Inception v4	86.23±0.60
ResNet34	86.68±1.30
ResNet101	88.47±1.50
ResNet50	**90.17±0.90**

Note: Performance of image classification of the CNN models for the dataset of grape leaves is expressed as a percentage±standard deviation, with the best result in bold type.

The pre-training model was tried to use on the imagenet[27] for transfer learning to initialize weights in ResNet50 and compared with the same network training from Xavier initialization[28]. The training and validation process of two kinds of weight initialization methods are presented in Figure 5. The convergence trend of the two networks was basically the same in training, but transfer learning model fluctuated more greatly in validation. The model using transfer learning was pretty much the same as the model using Xavier initialization and the dataset of this study to train from scratch. Enough data of the leaf dataset after data enhancement caused this situation where imagenet pre-training did not improve final classification accuracy, so Xavier initialization was used to initialize weights in CNN when the classifiers in this study were trained.

3.2 Accuracies and losses during the training of the classifiers

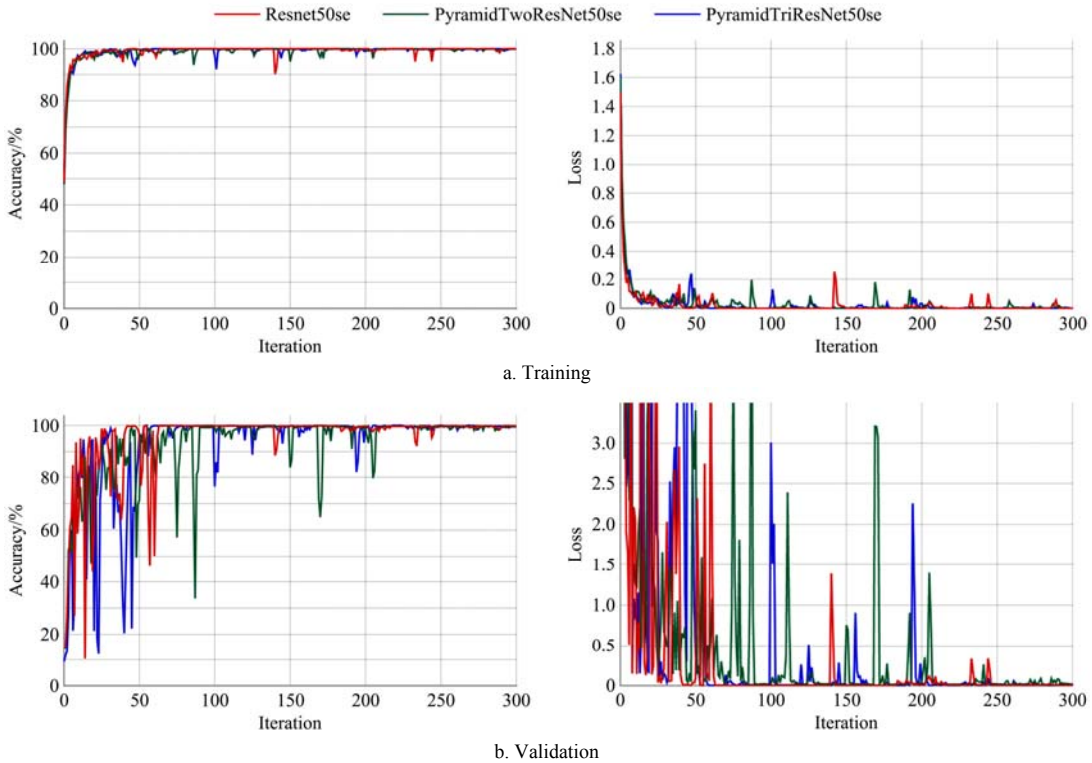
The accuracy and loss of training and validation of three classifiers are presented in Figure 6. The trend of convergence of the training loss and the consistent trend between training accuracy and validation accuracy indicated no obvious overfitting of the classifier. The validation of these models had some outliers, so the model in which the validation loss was reduced for the last time during training was used when the test data set was used to evaluate the classifier, i.e., the model at the 300th iteration was not necessarily used but was more likely to be the model with the lowest loss between iterations 200 and 300.



a. Training

b. Validation

Figure 5 Transfer learning using imagenet pre-training and Xavier initialized training model on ResNet50



a. Training

b. Validation

Figure 6 Training and validation accuracies and losses of three classifiers

3.3 Performance of the pyramid residual CNN classifiers

The performance of the CNN classification models with different structures for the 11 grape cultivars is listed in Table 3. All results presented in this section were the average results of 5-fold cross-validation. Among the classifiers, ResNet50 with an SE block attention mechanism performed better than models with VGG16 as the backbone. The accuracy of PyramidTwoResNet50se combined with two feature maps improved by 0.74%, while the accuracy of PyramidTriResNet50se with three feature maps improved by 2.01%. The PyramidTriResNet50 model performed best, with an average accuracy of fine-grained classification of 92.26%. These increases in accuracy indicated that the added feature-fusion mechanism extracted features more effectively and CNNs with residual structures performed better than sequentially connected CNNs.

The confusion matrix of the PyramidTriResNet50 model for the grape dataset is presented in Figure 6. The foci of the models could be determined by analyzing the Grad-Cams of some blades. Brighter regions indicated that the model paid more attention to these regions. The image of the training in Figure 7 is Hutai 8,

and the image of the test is Moon Drops. The ResNet50se backbone model was distracted by noise. Combined with one feature map, the PyramidTwoResNet50se model paid more attention to the central area of the blade. Combined with two feature maps, the attention of the PyramidTriResNet50se model was more focused on the edges and central region of the blade, indicating that the pyramid residual model could pay more attention to the important regions such as the edge and center of the leaf than the backbone model.

Cultivars 4 “Meili” and 5 “Italy” were more likely to be misclassified (Figure 8, Table 4), because their leaves are very similar in appearance, so classifying them using the depth CNN from the leaf image is difficult. This part of the misclassified images caused trouble for all classification models. The pyramid residual CNN for most of the cultivars could extract the deep features based on information for the shape, size, color, and texture from the leaf image without human intervention.

Table 3 Comparison of the pyramid residual CNN model and other models for identifying grape leaves

Model	Accuracy/%
VGG16+CosFace[14]	75.12±0.70
VGG16+Combined classifier[14]	76.04±1.70
VGG16+ArcFace[14]	81.32±1.20
ResNet50	90.06±1.00
ResNet50se	90.25±0.60
PyramidTwoResNet50se	90.78±0.90
PyramidTriResNet50se	**92.30±1.40**

Note: Comparison of the pyramid residual CNN model and other models for identifying grape leaves expressed as a percentage±standard deviation, with the best result in bold type.

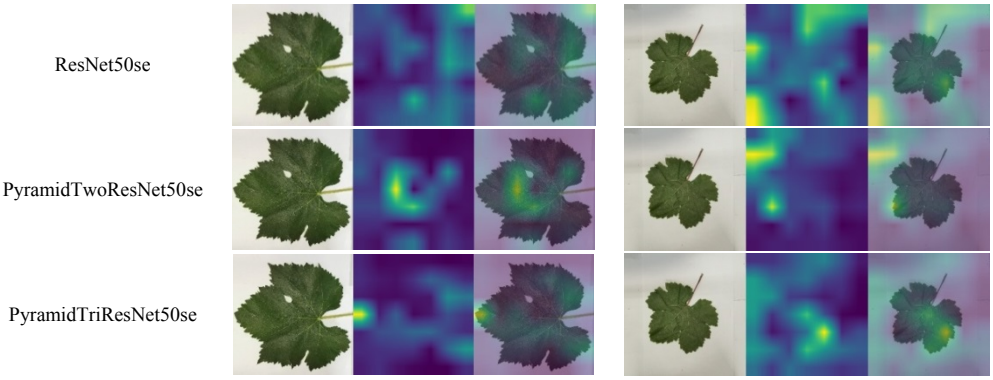
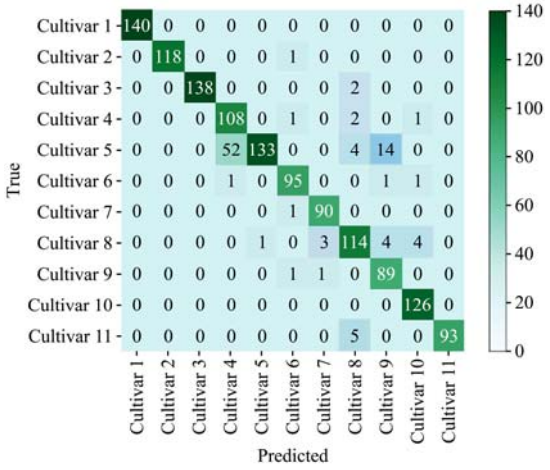


Figure 7 Grad-Cams of the backbone classifier and the pyramid CNN classifiers



Note: Cultivars 1-11: Cabernet Sauvignon, Chardonnay, Ecolly, Meili, Italia, Shine-Muscat, Crimson Seedless, Ruiduhongyu, Hutai 8, Wink, and Moon Drops.

Figure 8 Confusion matrix of the PyramidTriResNet50 model for the grape leaf dataset

The performance of the CNN classification models with different structures for the PlantVillage dataset is listed in Table 5. The accuracy of plant species classification was much higher compared with the fine-grained classification only for grape cultivars. In the proposed method in this study, the center and contour features of leaves are extracted by convolution. These features are fused to strengthen the attention of the model to the regions with morphological characteristics of leaves. This study improved the reliability of plant classification of different cultivars and provides a method for the identification of different grape varieties. This method could also be applied to other agricultural image classification and fine-grained classification problems.

Table 4 Accuracy of the classification of grape leaves for the 11 cultivars with PyramidTriResNet50se

True/Predicted	Precision/%	Recall/%	F1-score/%
Cultivar 1*	100	100	100
Cultivar 2	100	99.16	99.58
Cultivar 3	100	98.57	99.28
Cultivar 4	67.08	96.43	79.12
Cultivar 5	99.25	65.52	78.93
Cultivar 6	95.96	96.94	96.45
Cultivar 7	95.74	98.90	97.30
Cultivar 8	89.76	90.48	90.12
Cultivar 9	82.40	97.80	89.45
Cultivar 10	95.45	100	97.67
Cultivar 11	100	94.90	97.38

Note: * Cultivar 1-11 represent cultivars Cabernet Sauvignon, Chardonnay, Ecolly, Meili, Italia, Shine-Muscat, Crimson Seedless, Ruiduhongyu, Hutai 8, Wink, and Moon Drops.

Table 5 Comparison of the pyramid residual CNN model and other models for the PlantVillage dataset

Model	Accuracy/%
VGG16+ArcFace[14]	96.78±0.40
VGG16+Combined classifier[14]	96.85±0.40
VGG16+CosFace[14]	97.29±0.50
ResNet50	99.17±0.20
ResNet50se	99.42±0.10
PyramidTwoResNet50se	99.25±0.20
PyramidTriResNet50se	**99.46±0.10**

Note: Comparison of the pyramid residual CNN model and other models for the PlantVillage dataset expressed as a percentage±standard deviation, with the best result in bold type.

4 Conclusions

In this study, a pyramid residual CNN was developed for classifying the leaves of 11 grape cultivars. Two low-resolution feature images of the leaves were extracted firstly using two convolution layers with convolution cores of 3×3 and a stride of 2. The original image and the extracted feature map were then entered into three ResNet50se models, and the shape of the output convolution was adjusted by modifying the stride of some convolution layers in the network. These outputs of the same shape were fused by adding convolution kernels so that CNN could pay attention to the features extracted by different networks. The study showed that, compared with single-scale, a CNN with multi-scale input could improve the accuracy of leaf image classification. This study, however, also had two main problems:

1) The video memory space was larger than that of a single network due to the use of multiple networks with pyramid structures. This problem, however, should be alleviated with the development of hardware;

2) Some misclassification remained due to the similar morphologies of the cultivar leaves. The feature map in future research should be replaced by images of other plant organs as auxiliary features to improve the accuracy of classification by concatenating feature maps and fusing them by addition.

This method can carry out fine-grained classification better than existing methods and identify the leaves of different grape cultivars with similar morphologies, and offers higher accuracy. This method is a general classifier, so it can be applied to other problems of agricultural image classification and fine-grained classification. Future studies will reduce the background information of leaves in the field condition.

Acknowledgements

This work was financially supported by the National Key Research and Development Project (Grant No. 2020YFD1100601). The authors acknowledge Xin Yang for his help in making the grape dataset and also acknowledge William Blackhall for the language edition.

[References]

[1] FAO statistical database. Faostat. 2019. Available: https://www.fao.org/faostat/. Accessed on [2021-04-25].

[2] Macleod N, Benfield M, Culverhouse P. Time to automate identification. Nature, 2010; 467(7312): 154–155.

[3] Yousefi E, Baleghi Y, Sakhaei S M. Rotation invariant wavelet descriptors, a new set of features to enhance plant leaves classification. Computers and Electronics in Agriculture, 2017; 140: 70–76.

[4] Wu S G, Bao F S, Xu E Y, Wang Y X, Chang Y F, Xiang Q L. A leaf recognition algorithm for plant classification using probabilistic neural network. In: Proceedings of the 2007 IEEE International Symposium on Signal Processing and Information Technology, Giza, Egypt: IEEE, 2007; pp.11–16. doi: 10.1109/ISSPIT.2007.4458016.

[5] Saleem G, Akhtar M, Ahmed N, Qureshi W S. Automated analysis of visual leaf shape features for plant classification. Computers and Electronics in Agriculture, 2019; 157: 270–280.

[6] Xue J R, Fuentes S, Poblete-Echeverría C, Viljo C G, Tongson E, Du H J, et al. Automated Chinese medicinal plants classification based on machine learning using leaf morpho-colorimetry, fractal dimension and visible/near infrared spectroscopy. Int J Agric & Biol Eng, 2019; 12(2): 123–131.

[7] Wang B, Gao Y S, Yuan X H, Xiong S W, Feng X Z. From species to cultivar: Soybean cultivar recognition using joint leaf image patterns by multiscale sliding chord matching. Biosystems Engineering, 2020; 194: 99–111.

[8] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 2012; 60(6): 84–90. doi: 10.1145/3065386.

[9] He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the Computer Vision and Pattern Recognition (CVPR), Las Vegas: IEEE, 2016; pp.770–778. doi: 10.1109/CVPR.2016.90.

[10] Hall D, Mccool C, Dayoub F, Sunderhauf N, Upcroft B. Evaluation of features for leaf classification in challenging conditions. In: Proceedings of the Workshop on Applications of Computer Vision, Waikoloa, USA: IEEE, 2015; pp.797–804. doi: 10.1109/WACV.2015.111.

[11] Pereira C, Morais R, Reis M. Deep learning techniques for grape plant species identification in natural images. Sensors, 2019; 19(22): 4850. doi: 10.3390/s19224850.

[12] Yang H-W, Hsu H-C, Yang C-K, Tsai M-J, Kuo Y-F. Differentiating between morphologically similar species in genus *Cinnamomum* (Lauraceae) using deep convolutional neural networks. Computers and Electronics in Agriculture, 2019; 162: 739–748.

[13] Kaya A, Keceli A S, Catal C, Yalic H Y, Temucin H, Tekinerdogan B. Analysis of transfer learning for deep neural network based plant classification models. Computers and Electronics in Agriculture, 2019; 158: 20–29.

[14] Tavakoli H, Alirezazadeh P, Hedayatipour A, Banijamali Nasib A H, Landwehr N. Leaf image-based classification of some common bean cultivars using discriminative convolutional neural networks. Computers and Electronics in Agriculture, 2021; 181: 105935. doi: 10.1016/j.compag.2020.105935.

[15] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the Computer Vision and Pattern Recognition, 2014.

[16] Wang F, Xiang X, Cheng J, Yuille A L. NormFace: L_2 hypersphere embedding for face verification. In: Proceedings of the 25th ACM International Conference on Multimedia, 2017; pp.1041–1049. doi: 10.1145/3123266.3123359.

[17] Wang H, Wang Y T, Zhou Z, Ji X, Gong D H, Zhou J C, et al. CosFace: Large margin cosine loss for deep face recognition. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA: IEEE, 2018; pp.5265–5274. doi: 10.1109/CVPR.2018.00552.

[18] Deng J K, Guo J, Xue N N, Zafeiriou S. ArcFace: Additive angular margin loss for deep face recognition. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, USA: IEEE, 2019; pp.4685–4694. doi: 10.1109/CVPR.2019.00482.

[19] Lin T-Y, Dollar P, Girshick R, He K M, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the Computer Vision and Pattern Recognition, Honolulu, USA: IEEE, 2017; pp. 936–944. doi: 10.1109/CVPR.2017.106.

[20] Meng Y, Lin C-C, Panda R, Sattigeri P, Karlinsky L, Oliva A, et al. AR-Net: Adaptive frame resolution for efficient action recognition. In: Proceedings of the European Cofference on Computer Vision (2020ECCV), Springer, 2020; pp.86–104. doi: 10.1007/978-3-030-58571-6_6.

[21] Hughes D P, Salathé M. An open access repository of images on plant health to enable the development of mobile disease diagnostics through machine learning and crowdsourcing. 2015. arXiv: 1511.08060.

[22] International Plant Genetic Resources Institute (IPGRI), International Union for the Protection of New Varieties of Plants (UPOV), Office International de la Vigne et du Vin (OIV). Descriptors for grapevine (*Vitis* spp.). IPGRI, UPOV, OIV, 1997; 62p.

[23] Xu G Q, Li C, Wang Q. Unified multi-scale method for fast leaf classification and retrieval using geometric information. IET Image Processing, 2019; 13(12): 2328–2334.

[24] 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 2017 IEEE International Conference on Computer Vision (ICCV), Venice: IEEE, 2017; pp.618–626. doi: 10.1109/ICCV.2017.74.

[25] Hu J, Shen L, Albanie S, Sun G, Wu E H . Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018; 42(8): 2011–2023.

[26] Abadi M, Barham P, Chen J M, Chen Z F, Davis A, Dean J. TensorFlow: A system for large-scale machine learning. In: Proceedings of the 12th USENIX conference on Operating Systems and implementation, 2016; 265–283.

[27] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 2015; 115: 211–252.

[28] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. Journal of Machine Learning Research-Proceedings Track, 2010; 9: 249–256.