Plant Leaf Diseases Fine-grained Categorization using Convolutional Neural Networks

Yang Wu\textsuperscript{1,2}, Xian Feng\textsuperscript{1} and Guojun Chen\textsuperscript{2}

\textsuperscript{1}School of Intelligent Equipment Engineering, Wuxi Taihu University, Qianrong Road, No.68, Wuxi Jiangsu 214064, China
\textsuperscript{2}Jiangsu Key Lab of IoT Application Technology, Wuxi Taihu University, Wuxi Jiangsu 214064, China

Corresponding author: Guojun Chen (e-mail: wuy1@wxu.edu.cn).

This research was funded by General Project of Natural Science Research in Universities of Jiangsu Province (No. 20KJB210012), Outstanding Young Backbone Teacher of Jiangsu “Qinglan Project” (Su Teacher [2021] No.11).

ABSTRACT The difference of leaf disease images within the class is large but the difference between the class is small, so it is very important to represent the features of the local area of the target. Moreover, the complex network occupies a large amount of computer memory and wastes a large amount of computing resources, which is difficult to meet the needs of low-cost terminals. This paper proposes a fine-grained disease categorization method based on attention network to solve the problem. In “Classification Model”, attention mechanism is used to increase identification ability. “Reconstruction-Generation Model” were added during training and the “Classification Model” have to pay more attention to differentiate areas to find differences instead of paying more attention to global features. And adversarial loss was applied to distinguish the generated image from the original image to suppress the noise introduced by the “Discrimination Model”. Due to the feature that “Reconstruction-Generation Model” and “Discrimination Model” are only used in training and do not participate in the operation of inference phase, which can not increase the complexity of the model. Compared with the traditional classification network, the method of generalization ability enhancement further enhances the identification accuracy. And the method needs less memory but can achieve low performance terminal real-time identification of peach and tomato leaf diseases. And it can be applied in other crop disease identification fields with the similar application scenarios.

INDEX TERMS fine-grained categorization, attention, adversarial loss, reconstruction, generation

I. INTRODUCTION

Fine-grained image analysis (FGIA) has attracted much interest in computer vision and machine learning in recent years. The goal of FGIA in computer vision is to retrieve and identify images belonging to multiple subcategories of supercategories (i.e., meta-categories or basic categories), such as different kinds of animals, plants, different kinds of autos, different kinds of retail products, etc. The key challenge is to understand fine-grained visual differences that are sufficient to distinguish objects with highly similar overall appearance but different fine-grained characteristics. FGIA focuses on dealing with objects belonging to multiple subcategories in a meta-category. Usually, two species have very similar features but there are still many intra-category differences due to different postures, backgrounds, and camera angles. Therefore, in order to successfully classify the two species in fine granularity, the most important thing is to find the discriminative part in the image that can distinguish these two species to better represent the features of these distinctive regions. As deep-learning has become a powerful discriminant feature learning method and made significant breakthroughs in the field of FGIA, FGIA based on deep learning has greatly promoted the practical deployment of these methods in various fields [1-4]. Fine-grained image categorization, also known as subcategory recognition, identifies subcategories under a large category. Discriminative parts in fine-grained image categorization usually occupy only a small area of the image, which is a very popular research topic in computer vision, pattern recognition and other fields in recent years. Objects in fine-grained tasks typically share small inter-
class variances and large intra-class variances, as well as multiple object scopes and complex backgrounds, which makes the problem more complicated. The general image classification aims to separate the two objects. It can be seen that the two are easily distinguished from each other in terms of appearance and shape. The purpose of fine-grained image classification is to classify the same basic category of images (cars, dogs, flowers, birds, etc.) into more detailed subcategories. Compared with the general image task, the difference and difficulty lie in the finer granularity of the image category. Not only for computers, but also for the public, the difficulty and challenges of fine-grained image task are undoubtedly greater. Due to the small granularity of classification, fine-grained image classification is very difficult and even it is difficult for the experts in some categories to distinguish. Therefore, in order to successfully classify these two species which are very similar in fine granularity, the most important thing is to find the discriminative part in the image which can distinguish these two species, and to better represent the features of these discriminative regions.

The rest of the paper is organized as follows: Section 2 introduces the related work. Section 3 introduces the data and model for plant leaf disease images in detail and the detailed structure of the model. In section 4, the experiment is described and the results are analyzed. Finally, section 5 summarizes the article.

II. RELATED WORK

In the traditional image classification network, no matter how large the proportion of the important discriminant region in the image is in the whole image, the feature is extracted equally for the whole image. Therefore, if the same feature extraction and processing are carried out for some images with a small proportion in some discriminant areas, a large amount of background information that is not concerned will be trained, which increases the difficulty of image classification and reduces the accuracy of classification. The concept of fine-grained image categorization is introduced to solve such problems by focusing on small differences in images to achieve more accurate image classification, such as the classification of finer varieties of vegetables and fruits, and the classification of crop pests and diseases. At present, the conventional method of fine-grained image categorization is to obtain the region of the target, and then carry out fine classification of the target so that the network can better understand the classified object. Since deep convolutional networks can learn very robust image feature representation, most methods of fine-grained image categorization are based on deep convolutional networks. These methods can be roughly divided into the following four directions.

A. FINE-TUNING METHODS BASED ON CONVENTIONAL NEURAL NETWORK

This method uses universal and common deep convolutional neural network (DCNN) for fine-grained categorization, which is difficult to capture differentiated local details, so it is not commonly used now. For example, AlexNet [5], VGG [6], GoogleNet [7], ResNet [8] and DenseNet [9]. Since these networks have strong ability of feature representation, they can achieve better results in conventional image classification. However, in fine-grained categorization, the differences between different species are very subtle. Therefore, it is not ideal to directly apply the conventional image classification network to fine-grained image categorization. Inspired by the theory of transfer learning [10], one method aims to transfer the network trained on large-scale data to a fine-grained categorization and recognition task. The common solution is to use the network weights pre-trained on ImageNet as the initial weights, and then finetune the network weights on the fine-grained categorization dataset to get the final classification network. In [11], Zhang et al. further introduced the metric loss function into fine-tuning of the fine classification network. Specifically, three samples (Positive, Reference and Negative) are input into three networks with shared weights each time, and then feature outputs of the three networks are used to calculate the loss function. In addition to the traditional softmax loss function, the three feature outputs also constitute the generalized triplet loss.

B. METHODS BASED ON PART DETECTION

When humans distinguish similar objects, they usually find the areas with distinguishing characteristics through quick scanning, and then carefully compare and identify the areas. Similar to the human method, the target block-based detection method divides the fine-grained image categorization into two parts: discriminating region location and fine-grained feature learning in the region. Firstly, the location of the target is detected in the image, and then the part of the target with discriminative region is detected. Then, the target image (foreground) and the discriminative target region block are simultaneously sent into the deep convolutional network for feature extraction and classification. When positioning the discriminative region, Convolutional characteristic responses of deep neural networks are usually used in a strongly supervised or weakly supervised manner. In fine-grained feature learning, features are extracted from each region to be located, combined together and finally classified. Based on the method of object block detection, the bounding box annotations of targets and even the key feature points in the target image are often necessary in the training process. However, it is very difficult to obtain these annotations in practical applications. The representative one is the Part-RCNN method [12] proposed in 2014. The main idea is to use R-CNN [13], a classical method in target detection, to detect the position of the bird in the picture, including its head and its body, and then input the information of these three parts simultaneously into deep convolutional network.
for training. POOF [14] uses data mining techniques to learn a set of intermediate features, each of which is good at distinguishing between two specific categories based on the appearance of a particular part. In order to find the exact local area, Liu et al. [15] established a geometric appearance model based on the sample dog breed and its face, such as the dog’s face and eyes, to fine-classify specific categories of dogs. Yang et al. [16] proposed a template model for discovering common geometric patterns and co-occurrence statistics of patterns in object’s local area, and extracting features from aligned common patterns for fine-grained image categorization. The method of strong supervision requires not only category label, but also component label and key position box. This method has achieved good results, but the disadvantage is that it requires expensive manual labeling, and the location of manual labeling may not be the best discriminating area, which completely depends on the cognitive level of the annotator.

Strongly supervised fine-grained image categorization model, in model training, in order to obtain better classification accuracy, in addition to the category labels of images, additional manual annotations such as object bounding box and part annotations are also used. In recent years, many weakly supervised methods have been proposed [17-18]. This method uses attention mechanism, clustering and other means to automatically discover the different regions, and does not need the component labeling, only the classification label can complete the training. At present, there are many methods in this fields, from the perspective of the development trend. The fine-grained image categorization method based on object block detection has experienced a development process from manual feature engineering to multi-stage method and then to end to end.

C. METHODS BASED ON VISUAL ATTENTION MECHANISM

Vision attention mechanism is a special signal processing mechanism of human vision. Specifically, when the visual system is looking at something, it first scans the global image quickly to get the target area that needs attention, and then prevent other useless information from getting the target of interest. At present, the visual attention method based on CNN network is widely used in computer vision, including target detection, recognition and other tasks. In deep convolutional networks, attention model can also be used to find regions of interest or discriminative regions in images, and for different tasks, convolutional networks focus on different regions of interest. The basic idea behind the attention mechanism in computer vision is to teach the system to pay attention, ignoring irrelevant information and focusing on important information. The idea is to find correlations based on existing data, and then highlight some important features.

Since vision attention model-based methods can locate discriminating regions in images without additional annotation information, they have been widely used in fine-grained image categorization in recent years [19-21]. The representative work is a recurrent attention convolutional neural network (RA-CNN) proposed in 2017 [22]. This model imitates the region proposal network (RPN) in Faster-RCNN [23] and uses attention proposal network (APN) to locate the discriminative regions in the image. And rank loss function was used in the training process to ensure that the location of attention model is more effective each time.

In recent years, attentional mechanisms have also been applied to fine-grained image classification of crop diseases [24-25]. The NASNetLarge fine-grained classification model [24] based on the proposed attention mechanism that effectively utilizes the informative regions of an image, focus on the more discriminative regions of the image while improving the classification accuracy of the model in complex scenes, and describes the use of transfer learning to quickly construct several fine-grained image classification models of crop disease based on this attention mechanism. Lee et al. [25] develop a technique based on a Recurrent Neural Network (RNN) to automatically locate infected regions and extract relevant features for disease classification. RNN-based approach is more robust and has a greater ability to generalize to unseen infected crop species as well as to different plant disease domain images compared to classical CNN approaches.

D. METHODS BASED ON FINE-GRAINED FEATURE LEARNING

This method usually uses multiple DCNNs to discriminate similar features in fine-grained categorization. Lin et al. [26] proposed Bilinear convolutional neural network model (Bilinear CNN) to achieve a better representation of deep convolutional features. The bilinear model is a two-flow architecture consisting of two feature extractors, and the output is multiplied and combined at each location of the image using the cross product to obtain the image descriptor. Using VGG-D and VGG-M networks as reference networks, the cub200-2011 dataset achieves 84.1% classification accuracy without bounding Box annotations. With bounding Box, the classification accuracy is as high as 85.1%, and the architecture can model locally paired feature interactions in a translation-invariant manner, which is especially useful for fine-grained categorization. Wang et al. [27] proposed multiple granularity CNN, which consists of a set of parallel deep convolutional neural networks, each of which is optimized for classification in accordance with a given granularity. In other words, the multiple granularity CNN consists of a set of single-granularity descriptors. Saliency guidance in the hidden layer selects regions of interest (ROI) from a common pool of candidate image blocks generated from the bottom up. Learning discriminant feature representation from discriminant region plays a key role in fine-grained image categorization. As shown in Figure 1. (a) a fine-tuning method based on conventional image classification network. (b) Part Detection based on the target block. This kind of method firstly locates
the discriminant region, and then classified based on these regions. Most of these methods require annotation information on the object or local region, which is difficult to collect. (c) The method based on visual attention mechanism tries to automatically locate the discriminant region in an unsupervised way through the attention mechanism, so no additional labeling information is required. Such approaches often require additional network structures (such as attentional mechanisms), introducing additional computational overhead during the training and inference phases. (d) the fine-grained feature learning method shown in this paper.

![Image](47x247 to 288x327)

**FIGURE 1.** Four kinds of fine-grained image categorization methods.

**III. MATERIALS AND METHODS**

**A. DATA**

PlantVillage dataset [28] collects more than 50,000 images of 14 species of plants with 38 category labels. Among them, there are 2657 peach leaves in 2 categories, which are respectively healthy leaves and peach bacterial spot disease leaves, and 18162 tomato leaves in 10 categories, which are respectively healthy leaves and 9 kinds of disease leaves, are used as the basic dataset of crop disease images for the experiment, as shown in Figure 2. The above two types of fruits were divided into training-set, validation-set and testing-set according to 70%-10%-20%, respectively, for training model, selecting model and evaluating model performance.

![Image](47x513 to 287x633)

**FIGURE 2.** Examples of leaf diseases: Peach healthy (PH), Peach Bacterial spot (PBS), Tomato healthy (TH), Tomato Bacterial spot (TBS), Tomato Early blight (TEB), Tomato Late blight (TLB), Tomato Leaf mold (TLM), Tomato Mosaic virus (TMV), Tomato Septoria leaf spot (TSLS), Tomato Target spot (TTS), Tomato Two-spotted spider mite (TTSSM) and Tomato Yellow leaf curl virus (TYLCV), respectively.

**TABLE I**

| Class    | all  | training-set | validation-set | testing-set |
|----------|------|--------------|----------------|-------------|
| PH       | 360  | 252          | 36             | 72          |
| PBS      | 2297 | 1607         | 230            | 460         |
| Total    | 2657 | 1859         | 266            | 532         |

Table 1 and Table 2 respectively listed the number of samples of different diseases of the two fruits and the number of samples divided into different sample sets.

**TABLE II**

| Class | all  | training-set | validation-set | testing-set |
|-------|------|--------------|----------------|-------------|
| TH    | 1592 | 1115         | 159            | 318         |
| TBS   | 2127 | 1488         | 213            | 426         |
| TEB   | 1000 | 700          | 100            | 200         |
| TLB   | 1910 | 1337         | 191            | 382         |
| TLM   | 952  | 667          | 95             | 190         |
| TMV   | 373  | 262          | 37             | 74          |
| TSL   | 1771 | 1239         | 178            | 354         |
| TTS   | 1404 | 982          | 141            | 281         |
| TTSSM | 1676 | 1173         | 168            | 335         |
| TYLCV | 5357 | 3749         | 536            | 1072        |
| Total | 18162| 12713        | 1816           | 3633        |

**B. IDENTIFICATION MODELS FOR PLANT LEAF IMAGE DISEASES**

Disease is one of the main factors limiting crop cultivation. Crop infestation by diseases can lead to a significant reduction in agricultural yields, resulting in huge losses to the agricultural economy. Therefore, early identification of diseases is critical to the selection of the correct treatment, as well as an important prerequisite for reducing crop losses and pesticide use [29-30]. On the one hand, the disease affects the yield and quality; on the other hand, excessive chemical control results in drug residues and brings public hazards. With the improvement of people's living standards, the demand for fruits and other crops is increasing. Early diagnosis and treatment are problems that must be solved. With the rapid development of computer technology, the research of crop disease identification using computer vision technology appeared [31-32]. In recent years, the classification and detection of plant disease leaves using deep convolutional neural network can achieve satisfactory results [33-36]. In the recognition of crop leaf disease in agriculture, there are some great similarities between different kinds of disease images of leaves, and great differences between the same kind of disease images. The fact is that the difference between different kinds of disease images is small while the difference within the same kind of disease images is large, which increases the difficulties to the actual recognition, which is a typical fine-grained categorization problem. Many agricultural fields are data vacuum areas, and the difficulty of disease identification lies in the collection and annotation of a large number of image data, which requires manual data collection in the field, which is time-consuming and laborious. At the same time, it is also difficult to accurately annotate the collected data. Due to the lack of experience, it is difficult to judge whether the identification is accurate. Experienced experts are required to accurately annotate the data. In order to adapt the training model to the characteristics of specific data in the agricultural field, this paper proposes a fine-grained disease recognition method based on attentional depth neural network, as shown in Figure 1(d). Attention mechanism is used in the classification process to focus on discriminative features.
network to increase the discriminant ability, and the “Reconstruction Network” (R) and “Generation Network” (G) is added during training, forcing the classification network to learn the discriminant parts of the image rather than focusing on the global features. Since the “Reconstruction and Generation” method is only used in training and does not increase the complexity of the network, so it does not need any additional knowledge during identification. In addition to standard classification network feedforward stage of the model, the stage of reasoning does not participate in the operation. In the actual deployment and production environment, the model will not introduce additional computational overhead and storage consumption.

FIGURE 3. Fine-grained disease categorization method based on attentional deep neural network.

A fine-grained disease categorization method based on attentional deep neural network is proposed. Reconstruction and generation learning method is adopted to enhance the difficulty of fine-grained categorization and train classification models to acquire expert knowledge. Model for plant leaf disease identification is composed of three parts. As shown in Figure 3, the red box is "Identification Model", which is used both in training and inference phase. While the blue box and the green box are “Discrimination Model” and “Reconstruction & Generation Model” respectively, which are only used in training. The “Identification Model” is mainly composed of “Encoder Network”, “Attention Network” and “Classification Network”. The “Encoder Network” is used to extract the input image features.

“Reconstruction & Generation Model” is used to force the network to learn the distinguishing parts of an image, instead of paying more attention to the global features, focusing on the details of classification, rather than the whole. The method of breaking first and generating later is adopted. Specifically, for “reconstruction”, the author first divides the input image into many local regions and reassembles them. To correctly identify these corrupted images, the classification network must focus more on discriminative information.

In the inference stage, the model with the stage of reasoning does not participate in the operation. In the actual deployment and production environment, the model will not introduce additional computational overhead and storage consumption. Compared with the traditional classification network, the generalization ability of the proposed method is enhanced, and the recognition accuracy is further improved.

C. COMPONENT MODULE OF EACH MODEL

The “Identification Model” is composed of “Encoder Network” (E), “Attention Network” (A) and “Classification Network” (C). “E” is used to extract the input image features, and two vectors with length of 256 are obtained. Then through the technique of heavy parameter reparemerization, the hidden vector is obtained. The hidden vector is not a simple reconstruction of the original data, but also introduces noise and increases the generalization ability of the model.

After that, the classification network determines the category. In the inference stage, is directly replaced by , instead of using reparemerization technique to generate . The detailed “Identification Model” is shown in Figure 4 and the sizes of every layer are shown in Table 3.

The Encoder converts an input image of size into two vectors of mean and variance of size 256. The detailed “Encoder Network” is shown in Figure 4 and output sizes of every layer are shown in Table 1, “E” consists of a series of convolution layers. It is composed of Conv, layers, Scale, Reduce mean, Scale fc and FC. layers is made up of four alternating Scale, SENet and Downsample, and Scale is the ResNet module, which is used to extract features. SENet is shown as Figure 5. Downsample is used to decrease the size of each feather map and increase the number of channels, after each layer, the number of channels is doubled and the size is halved. The input of the model is a image, the size of the input vector is changed to after Conv layer, after layers, the size is . Reducemean is global pooling, and the structure of Scale_fc is shown in Figure 4 for better access to global information.
FIGURE 4. Encoder Network and Attention Network.

TABLE III
OUTPUT SIZE OF THE LAYER IN THE ENCODER AND ATTENTION NETWORK

| Layer             | Input Size       | Output Size       |
|-------------------|------------------|-------------------|
| Conv (3 × 3, 16)  | 128 × 128 × 3    | 128 × 128 × 16    |
| Scale 0           | 128 × 128 × 16   | 128 × 128 × 16    |
| SENet             | 128 × 128 × 16   | 128 × 128 × 16    |
| Downsample 0      | 128 × 128 × 16   | 64 × 64 × 32      |
| ...               | ...              | ...               |
| Downsample 3      | 16 × 16 × 128    | 8 × 8 × 256       |
| Scale block       | 8 × 8 × 256      | 8 × 8 × 256       |
| Reduce mean       | 8 × 8 × 256      | 256               |
| Scale fc          | 256              | 256               |

FIGURE 5. “SENet” and “Classification Network”.

“Reconstruction and Generation Model” is mainly composed of “Encoder Network” (E), “Attention Network” (A) and “Reconstruction Network” (R) and “Generation Network” (G). “R” is a fixed operation process which does not need training, to disturb the spatial distribution of the local area, the original input image block, after the input image is divided into many of the local area, they were then shuffled through a recombination mechanism and recombined into a single image. Details are as follows in Figure 6.

An image with a size of 128×128 is divided into N×N subregions (N is 8), and each region is marked \( R_{i,j} \), i and j represent row and column indexes respectively, 1 ≤ i,j ≤ N, to scramble these local regions within the plane region. Specifically, a random vector \( q_r \) is generated for the j-th row of r, where the value of the i-th element is \( q_r \), i=r+r, where r is a random variable in the range of \([-k, k]\) and obeys the uniform distribution of \( U(-k, k) \). k is an adjustable parameter, 1 ≤ k ≤ N, which defines the adjacent range of the transformable region, here, take 2. A new arrangement of the regions in the j-th row can then be obtained by sorting the array \( q_r \), which destroys the global structure and ensures that local regions are recombined in an adjustable range. Since

the global structure is broken, in order to recognize these randomly shuffled images, the classification network must seek discriminant regions and then learn the subtle variations of these categories. In practical application, rearranging the image first and then rearranging the rows. After dividing the image into N columns, exchanging the columns in the same row first, and then exchanging the data in the rows in the same column after all the rows are exchanged. Then rearranging the images using the rearranged array. In order to reconstruct the image, a generation network is used to try to restore the original spatial layout of the local areas to simulate the semantic correlation between the local areas.

“Generation Network” is used to convert vector to \( X \), restoring the latent vector \( z \) of size 256 to an image of size 128×128×3. The detailed “Generation Network” is shown in Figure 7(a) and the sizes of every layer are shown in Table 4. After 6 layers, the size is changed to 128×128×32, and after Conv, the size is changed to 128×128×3, which is the same size as the input image.

| Layer             | Input Size       | Output Size       |
|-------------------|------------------|-------------------|
| FC (4096)         | 256              | 4096              |
| Reshape (2, 1024) | 4096             | 2 × 2 × 1024      |
| Upsample 0        | 2 × 2 × 1024     | 4 × 4 × 512       |
| Scale 0           | 4 × 4 × 512      | 4 × 4 × 512       |
| ...               | ...              | ...               |
| Upsample 5        | 64 × 64 × 32     | 128 × 128 × 16    |
| Scale 5           | 128 × 128 × 16   | 128 × 128 × 16    |
| Conv (3 × 3, 3)   | 128 × 128 × 16   | 128 × 128 × 3     |

“Discrimination Model” is mainly composed of “Encoder Network” (E), “Attention Network” (A) and “Discriminator Network” (D). “D” will be able to distinguish between the generated and realistic images as far as possible, so the score...
for the original image shall be as high as possible, and the scores for the generated images shall be as low as possible. The detailed is shown in Figure 7(b).

**D. LOSS FUNCTION**

Model for plant leaf disease identification is composed of three parts. All three parts are involved in back propagation and forward reasoning during training. Therefore, Loss also consists of three parts. Specifically, it hopes to minimize the following objectives.

$$L = L_{cls} + L_{rec} + L_{dis}$$

In the model, the goal of the “Identification Model” is to minimize cross entropy loss $L_{cls}$, which represents classification accuracy. “Reconstruction and Generation model” is to keep an image as original as possible. Minimize $L_{rec}$ to reduce the introduced noise, so that the network can rebuild the damaged image and increase the semantic correlation between local information. The goal of the discriminator model is to try to differentiate the reconstructed and realistic images, which is adversarial loss. The training pipeline of proposed algorithm is as follows.

**Algorithm:**

- Initial parameters of the models: $\theta_e, \theta_r, \theta_g, \theta_d$
- While training do
  - Identification model:
    $$x^{\text{real}} \leftarrow \text{batch of images sampled from the dataset}.$$  
    $$z^{\text{real}} \leftarrow E_{\theta_e}(x^{\text{real}}).$$  
    $$z^{\text{real}} = \mu^{\text{real}} + \varepsilon \sigma^{\text{real}}$$  
    with $\varepsilon \sim N(0, I_d)$
    $$C_{\text{label}} \leftarrow C_{\theta_e}(z^{\text{real}}).$$

- Reconstruction & Generation model:
  $$x^{\text{rec}} \leftarrow R(z^{\text{real}}).$$  
  $$z^{\mu^{\text{rec}}} \leftarrow E_{\theta_r}(x^{\text{rec}}).$$  
  $$z^{\mu^{\text{rec}}} = \mu^{\text{rec}} + \varepsilon \sigma^{\text{rec}}$$  
  with $\varepsilon \sim N(0, I_d)$
  $$C_{\text{label}} \leftarrow C_{\theta_r}(z^{\text{rec}}).$$

- Discrimination model:
  $$I_{\text{label}} \leftarrow D_{\theta_d}(z^{\text{rec}}).$$  
  $$I_{\text{label}} \leftarrow D_{\theta_d}(z^{\text{real}}).$$

[Compute losses gradients and update parameters.]

$$\theta_e \leftarrow KL(P(z^{\text{real}}), P(z))$$  
$$\theta_r \leftarrow C_{\theta_r}(z^{\text{rec}}) - D_{\theta_d}(x^{\text{rec}})$$  
$$\theta_g \leftarrow C_{\theta_g}(z^{\text{rec}}) - D_{\theta_d}(x^{\text{rec}})$$  
$$\theta_d \leftarrow D_{\theta_d}(z^{\text{rec}}) - D_{\theta_d}(z^{\text{real}})$$

**VOLUME XX, 2017**

9

---

### IV. EXPERIMENTS AND RESULTS

**A. EXPERIMENTAL SETUP**

The experimental configuration environment of this paper is as follows: Ubuntu16.04 LST 64-bit system, processor Intel Core i5-11400, memory is 16 GB, graphics card is GeForce GTX1080ti (11G), using Tensorflow-GPU1.4 deep learning framework and “python” programming language.

**B. PERFORMANCE COMPARISON**

Only the network model used in the training process is changed, and the proposed model is compared with the classical convolutional neural network. In the experiment, the classification accuracy of the testing-set was used as the main evaluation index of the experimental results. The higher the classification accuracy, the better the performance of the model, the more outstanding the generalization ability. Table 5 lists the comparison of identification indexes of peach leaf diseases under different neural network models, and Table 6 lists the comparison of identification indexes of tomato leaf diseases under different neural network models. The comparison experiments were all completed under the same conditions.

**TABLE V**

Comparison of identification indexes of peach leaf diseases in different networks

| Class | ResNe t-50 | MobileNet-V2 | Biline ar CNN | DCL | PMG | Proposed |
|-------|------------|-------------|---------------|-----|-----|----------|
| PH    | 91.53      | 94.23       | 94.92         | 94.68 | 95.18 | 95.09    |
| BBS   | 92.33      | 94.84       | 94.82         | 94.98 | 94.98 | 95.54    |
| all   | 92.22      | 94.76       | 94.83         | 94.94 | 95.01 | 95.48    |

**TABLE VI**

Comparison of identification indexes of tomato leaf diseases in different networks

| Class | ResNe t-V2 | MobileNet-V2 | Biline ar CNN | DCL | PMG | Proposed |
|-------|------------|-------------|---------------|-----|-----|----------|
| TH    | 90.22      | 93.65       | 94.39         | 93.73 | 94.59 | 94.62    |
| TBS   | 91.06      | 92.77       | 94.02         | 93.39 | 93.69 | 95.06    |
| TEB   | 90.20      | 92.22       | 92.28         | 94.04 | 93.14 | 94.57    |
| TLB   | 90.11      | 91.99       | 93.28         | 94.04 | 93.48 | 93.28    |
| TLM   | 88.07      | 91.17       | 90.74         | 92.05 | 92.08 | 93.62    |
| TMV   | 88.32      | 92.57       | 91.44         | 91.82 | 93.49 | 93.72    |
| TSLS  | 90.30      | 92.83       | 94.01         | 94.34 | 93.75 | 93.46    |
| TTS   | 90.86      | 93.10       | 93.45         | 93.29 | 94.03 | 94.62    |
| TTSSM | 91.25      | 92.82       | 93.30         | 94.20 | 93.74 | 94.73    |
| TYLCV | 92.13      | 93.07       | 94.10         | 95.29 | 94.00 | 95.13    |
| all   | 90.87      | 92.85       | 93.34         | 94.04 | 93.78 | 94.35    |

As can be seen from Table 5 and Table 6, the recognition accuracy of fine-grained disease categorization method based on attentional depth neural network has been improved in the testing-set of various categories of two different crops, which reflects the effectiveness of using attentional mechanism and “reconstruction and generation” method to improve network performance. Comparing the improved convolutional neural network with several more advanced convolutional neural networks, including ResNet-50 [8], MobileNet-V2 [37], Bilinear CNN [26], DCL [38] with ResNet-50 and PMG [39] with ResNet-50, which use these networks to diagnose and identify peach and tomato diseased leaves. Table 5 and 6 list...
the classification accuracy of peach and tomato disease leaves with each different disease and all diseases.

C. ABLATION EXPERIMENT

The proposed fine-grained disease categorization model based on attentional depth neural network is composed of three parts. They are “Identification Model”, “Discrimination Model” and “Reconstruction & Generation Model” respectively. In order to verify the effectiveness of each model, the ablation experiment was performed. In Table 6 and 7, “i” is “Identification Model” without “Attention Network”. “a” stands for “Attention Model”. “i+a” is “Identification Model” include “Attention Network”. “d” stands for “Discrimination Model”. “i+d” stands for “Reconstruction & Generation Model”.

| Class  | i     | i+a   | i+d+r | i+a+d+r(Ours) |
|--------|-------|-------|-------|---------------|
| PH     | 92.94 | 93.82 | 94.45 | 95.09         |
| PBS    | 93.30 | 94.31 | 94.58 | 95.54         |
| all    | 93.25 | 94.24 | 94.56 | 95.48         |

| Class  | i     | i+a   | i+d+r | i+a+d+r(Ours) |
|--------|-------|-------|-------|---------------|
| TH     | 91.58 | 92.90 | 93.50 | 94.62         |
| TBS    | 91.83 | 92.53 | 93.44 | 95.06         |
| TEB    | 91.72 | 92.20 | 92.51 | 94.57         |
| TLB    | 91.21 | 92.22 | 91.70 | 94.58         |
| TLM    | 89.67 | 90.65 | 90.70 | 93.62         |
| TMV    | 89.68 | 91.35 | 91.06 | 93.72         |
| TLS    | 91.92 | 92.65 | 93.48 | 93.46         |
| TTS    | 91.68 | 92.67 | 93.08 | 94.62         |
| TTSSM  | 92.33 | 92.85 | 93.32 | 94.73         |
| TLYLCV | 93.31 | 93.47 | 94.22 | 95.13         |
| all    | 92.06 | 92.72 | 93.21 | 94.35         |

Table 7 and 8 lists the classification accuracy of peach and tomato disease leaves with each different disease and all diseases combined by different modules in our method. From the tables, we can see that all the models in our method are valid and significantly improve classification accuracy. And our combined method yielded the best results.

V. CONCLUSIONS

In deep learning research, in order to pursue the higher performance, it is necessary to study powerful and complex model networks. However, complex networks usually are limited by storage space and computing resources, which are difficult to be effectively applied to low-performance terminals. Moreover, complex networks have redundant parameters. The application of deep neural network to crop leaf disease identification has been widely studied in recent years. There is a great similarity among leaf disease images. The fact is that the difference within the classes is large but the difference between the classes is small, which brings great difficulty to the actual recognition. Therefore, exquisite feature representation of the local area of the target plays a key role in the fine-grained categorization task. Aiming at the deficiency of deep neural network in crop disease identification, a fine-grained disease identification method based on attentional deep neural network was proposed for peach and tomato disease leaf identification.

In order to learn discriminant regions and characteristics, add the “Reconstruction and Generation Model” during training, forcing the apart large part of the classification of network to learn the image, rather than paying more attention to the global characteristics. In order to correctly identify the damaged image, classification network must focus more on discriminant area to find differences. Therefore, attention mechanism is added on the basis of the general network classification model, and adversarial loss is used to distinguish the generated image from the original image to suppress the noise introduced by the destruction. Since the reconstruction and generation methods are only used in training and do not increase the complexity of the network, there isn’t any additional knowledge during identification. In addition to standard classification network feedforward stage, the model in the stage of reasoning does not participate in the operation. In the actual deployment and production environment, the model will not introduce additional computational overhead and storage consumption. Experimental results show that the categorization accuracy of the method can reach a high level.

Compared with the traditional classification network, the generalization ability of the proposed method is enhanced, and the categorization accuracy is further improved. In addition, the advantage of this model less memory, which makes it can run on low performance terminal. Integrate it into mobile applications for mobile devices is feasible, such as integrated into the smart phone, unmanned aerial vehicles and other automatic agricultural vehicles for growers or agronomists, implementation of large-scale cultivation of real-time monitoring and disease detection. The model can identify different diseases of plant leaves and meet the needs of tomato disease identification in agricultural production and it provides a theoretical basis for the development of intelligent identification device of plant leaf diseases, which can be extended to other similar application scenarios of crop disease categorization.

The future will mainly be carried out from the following aspects: (i) The detection of leaf diseases is also a problem that needs further study. Research on crop disease detection based on UAV was carried out, and the crop disease recognition algorithm was applied to UAV to collect images. (ii) If there is no large amount of annotated data as training samples to train the generated model, the generated model cannot be directly trained in this way, and the number of samples needs to be expanded first. In practice, it is difficult to collect images of diseased leaves, so the problem of few-shot learning is urgent to be solved. (iii) Research on early disease detection. The symptoms of early diagnosis are not obvious, and both visual observation and visual detection are difficult. However, the research significance and demand of
early diagnosis are greater, which is more conducive to preventing the spread of the disease. Therefore, early prediction and diagnostic models can be studied according to the imaging characteristics of different crop growth cycles, and an early warning mechanism can be established. In practical application, its significance will be greater than the detection and identification of disease in the middle and late period. To sum up, we will strive to achieve continuous improvement of performance and apply it to practical agricultural production.

REFERENCES

[1] X. Wei et al., “Fine-Grained Image Analysis with Deep Learning: A Survey,” IEEE Trans. Pattern Anal. Mach. Intell., to be published. DOI: 10.1109/TPAMI.2021.3126648.

[2] J. Yin, A. Wu and W. Zhang, “Fine-grained person re-identification,” Int. J. Comput. Vis., vol. 128, no. 6, pp. 1654–1672, Jan. 2020.

[3] S. D. Khan and H. Ullah, “A survey of advances in vision-based vehicle re-identification,” Comput. Vis. Image Und., vol. 182, pp. 50–63, May 2019.

[4] X. Wei, Q. Cui, L. Yang, P. Wang and L. Liu, “RPC: A large-scale retail product checkout dataset,” 2019, arXiv:1901.07249. [Online]. Available: http://arxiv.org/abs/1901.07249

[5] A. Kirzhnitsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Proces. Syst. (NIPS), Lake Tahoe, NV, USA, Dec. 2012, pp. 1097–1105.

[6] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556

[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going Deeper with Convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, Jun. 2015, pp. 1–9.

[8] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[9] G. Huang, Z. Liu, V. D. M. Laurens, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 2261–2269.

[10] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue and G. Zhang, “Transfer learning using computational intelligence: A survey,” Knowl-Based Syst., vol. 80, pp. 14–23, May 2015.

[11] X. Zhang, Z. Feng, Y. Lin, and S. Zhang, “Embedding label structures for fine-grained feature representation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 1114–1123.

[12] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, “Part-Based R-CNNs for Fine-Grained Category Detection,” in Proc. Euro. Conf. Comput. Vis. (ECCV), Zurich, Switzerland, Sep. 2014, pp. 834–849.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Columbus, OH, USA, Jun. 2014, pp. 580–587.

[14] T. Berg and P. N. Belhumeur, “POOF: Part-based one-vs-one features for fine-grained categorization, face verification, and attribute estimation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Portland, OR, USA, Jun. 2013, pp. 955–962.

[15] J. Liu, A. Kanazawa, D. Jacobs, and P. Belhumeur, “Dog breed classification using part localization,” in Proc. Euro. Conf. Comput. Vis. (ECCV), Florence, Italy, Oct. 2012, pp. 172–185.

[16] S. Yang, L. Bo, J. Wang, and L. Shapiro, “Unsupervised template learning for fine-grained object recognition,” in Proc. Adv. Neural Inf. Proces. Syst. (NIPS), Lake Tahoe, NV, USA, Dec. 2012, pp. 3122–3130.

[17] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, “The application of two-level attention models in deep convolutional neural network for fine-grained image classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, Jun. 2015, pp. 842–850.

[18] M. Simon, and E. Rodner, “Neural Activation Constellations: Unsupervised Part Model Discovery with Convolutional Networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Santiago, Chile, Dec. 2015, pp. 1143–1151.

[19] R. J. et al., “Attention Convolutional Binary Neural Tree for Fine-Grained Visual Categorization,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, Jun. 2020, pp. 10465–10474.

[20] L. Zhang, S. Huang, W. Liu, and D. Tao, “Learning a Mixture of Granularity-Specific Experts for Fine-Grained Categorization,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Seoul, Korea, Oct. 2019, pp. 8330–8339.

[21] F. Zhang, G. Zhai, M. Li, and Y. Liu, “Three-branch and Mutil-scale learning for Fine-Grained Image Recognition,” 2020, arXiv: 2003.09150. [Online]. Available: http://arxiv.org/abs/2003.09150

[22] J. Fu, H. Zhang, and T. Mei, “Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 4476–4484.

[23] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[24] G. Yang, Y. He, Y. Yang and B. Xu, “Fine-grained image classification for crop disease based on attention mechanism,” Front. Plant Sci., 11:854, Dec. 2020. DOI: 10.3389/fpls.2020.600854.

[25] S. H. Lee, H. Goeau, P. Bonnet, and A. Joly, “Attention-based recurrent neural network for plant disease classification,” Front. Plant Sci., 11: 601250, Dec. 2020. DOI: 10.3389/fpls.2020.601250.

[26] T. Lin, A. Roychowdhury, and S. Maji, “Bilinear CNN Models for Fine-Grained Visual Recognition,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Santiago, Chile, Dec. 2015, pp. 1449–1457.

[27] D. Wang, Z. Shen, J. Shao, X. Xue, and Z. Zhang, “Multiple granularity descriptors for fine-grained categorization,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Santiago, Chile, Dec. 2015, pp. 2399–2406.

[28] D. P. Hughes and M. Salathe, “An open access repository of images on plant health to enable the development of mobile disease diagnostics,” 2015, arXiv:1511.08060. [Online]. Available: http://arxiv.org/abs/1511.08060

[29] S. S. Chouhan, U. P. Singh, and S. Jain, “Applications of Computer Vision in Plant Pathology: A Survey,” Arch. Comput. Methods Eng., vol. 27, no. 2, pp. 611–632, Apr. 2020.
[30] Abade, P. A. Ferreira, and F. D. Vidal, “Plant diseases recognition on images using convolutional neural networks: A systematic review,” *Comput. Electron. Agric.*, vol. 185, pp. 75–105, 2021.

[31] S. S. Chouhan, U. P. Singh and S. Jain, “Automated Plant Leaf Disease Detection and Classification Using Fuzzy Based Function Network,” *Wireless Pers. Commun.*, vol. 121, no. 3, pp. 1757–1779, 2021.

[32] S. S. Chouhan, U. P. Singh, U. Sharma, and S. Jain, “Leaf disease segmentation and classification of Jatropha Curcas L. and Pongamia Pinnata L. biofuel plants using computer vision based approaches,” *Measurement*, vol. 171, pp. 594–603, 2023.

[33] S. M. Hassan, A. K. Maji, M. Jasinski, Z. Leonowicz, and E. Jasinska, “Identification of Plant-Leaf Diseases Using CNN and Transfer-Learning Approach,” *Electronics*, vol. 10, no. 12, 1388, 2021.

[34] J. Boulent, S. Foucher, J. Theau and P. Stcharles, “Convolutional Neural Networks for the Automatic Identification of Plant Diseases,” *Front. Plant Sci.*, 10:941, Jul. 2019. DOI: 10.3389/fpls.2019.00941.

[35] U. P. Singh, S. S. Chouhan, and S. Jain, “Multilayer Convolution Neural Network for the Classification of Mango Leaves Infected by Anthracnose Disease,” *IEEE Access*, vol. 7, pp. 43721–43729, 2019.

[36] A. M. Roy, R. Bose, and J. Bhaduri, “A fast accurate fine-grain object detection model based on YOLOv4 deep neural network,” *Neural Comput. Appl.*, to be published. DOI: 10.1007/s00521-021-06651-x.

[37] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 4510-4520.

[38] Y. Chen, Y. Bai, W. Zhang, and T. Mei, “Destruction and Construction Learning for Fine-Grained Image Recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 5152-5161.

[39] R. Du et al., “Fine-Grained Visual Classification via Progressive Multi-Granularity Training of Jigsaw Patches,” in *Proc. Euro. Conf. Comput. Vis. (ECCV)*, Glasgow, UK, Aug. 2020, pp. 153–168.

**YANG WU** was born in Shandong, China, in 1986. He is currently working as an associate professor of Wuxi Taihu University, and was supported as outstanding young backbone teacher of Jiangsu “Qinglan Project”. His current research interests include deep learning, machine vision, plant phenotyping.

**Xian Feng** received her M.S. degree from Nanjing University of Aeronautics and Astronautics. She is currently an associate professor of Wuxi Taihu University. Her current research interests include deep learning, big data and so on.

**GUOJUN CHEN** received the M.S. degree from Jiangnan University in 2009. He is currently an associate professor of Wuxi Taihu University. His current research interests include deep learning, machine vision, plant phenotyping.