frontiers in Plant Science

ORIGINAL RESEARCH
published: 22 December 2020
doi: 10.3389/fpls.2020.600854

Fine-Grained Image Classification for Crop Disease Based on Attention Mechanism

Guofeng Yang[1,2,3], *Yong He*[1]*, *Yong Yang*[2,3] *and Beibei Xu*[2,3]

[1] College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou, China, [2] Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing, China, [3] Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs, Beijing, China

Fine-grained image classification is a challenging task because of the difficulty in identifying discriminant features, it is not easy to find the subtle features that fully represent the object. In the fine-grained classification of crop disease, visual disturbances such as light, fog, overlap, and jitter are frequently encountered. To explore the influence of the features of crop leaf images on the classification results, a classification model should focus on the more discriminative regions of the image while improving the classification accuracy of the model in complex scenes. This paper proposes a novel attention mechanism that effectively utilizes the informative regions of an image, and describes the use of transfer learning to quickly construct several fine-grained image classification models of crop disease based on this attention mechanism. This study uses 58,200 crop leaf images as a dataset, including 14 different crops and 37 different categories of healthy/diseased crops. Among them, different diseases of the same crop have strong similarities. The NASNetLarge fine-grained classification model based on the proposed attention mechanism achieves the best classification effect, with an F_1 score of up to 93.05%. The results show that the proposed attention mechanism effectively improves the fine-grained classification of crop disease images.

Keywords: crop disease, fine-grained, image classification, attention mechanism, fine-tuning

OPEN ACCESS

Edited by:
Spyros Fountas, Agricultural University of Athens, Greece

Reviewed by:
Shitala Prasad, NTU Institute of Structural Biology (NISB), Singapore
Ning Yang, Jiangsu University, China

***Correspondence:**
Yong He
yhe@zju.edu.cn

Specialty section:
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

Received: *31 August 2020*
Accepted: *30 November 2020*
Published: *22 December 2020*

Citation:
Yang G, He Y, Yang Y and Xu B (2020) Fine-Grained Image Classification for Crop Disease Based on Attention Mechanism. Front. Plant Sci. 11:600854. doi: 10.3389/fpls.2020.600854

INTRODUCTION

Outbreaks of crop disease have a significant impact on the yield of agricultural production. Often, large-scale disease outbreaks destroy crops that have taken considerable efforts to grow, causing irreparable damage. Even without large-scale disease outbreaks, small-scale emergence can cause serious losses to crop yield and quality (Mutka and Bart, 2015). Therefore, developing techniques to accurately classify crop leaf disease categories is critical for disease prevention. With advances in image classification technology, researchers in the field of crop disease have gradually come to use deep learning approaches (Ramcharan et al., 2017; Fuentes et al., 2018; Liu B. et al., 2020). To date, research on the general classification of crop diseases has made several remarkable achievements in terms of better classification. However, for some fine-grained crop leaf diseases, there are still many difficulties.

Fine-grained image classification aims to classify sub-categories of a single larger category through fine-grained images (Peng et al., 2018). Examples include Stanford Cars (Yu et al., 2018; Tan and Le, 2019), CUB-200-2011 (Chen et al., 2019; Zhuang et al., 2020), FGVC Aircrafts

(Ding et al., 2019; Sun et al., 2020), and Oxford 102 Flowers (Dubey et al., 2018; Touvron et al., 2019). Fine-grained image classification models can be divided into algorithms based on strong supervision and algorithms based on weak supervision, which depends on how much supervision information can be used. For classification models based on strong supervision information, superior classification accuracy during model training requires artificial annotation information, such as object bounding boxes and part annotation, in addition to image-level category labels. Fine-grained image classification models based on weakly supervised information are similar, but also require the use of global and local information. Weakly supervised fine-grained classification attempts to achieve better local information capture without resorting to the key point information of object parts. As our goal is fine-grained image classification, we need to build a model that can identify the most discriminating image features. Therefore, it is vital to detect subtle discriminatory features from similar regions (Ou et al., 2016; Zhang et al., 2016). Because the occurrence of crop diseases is often not controlled by humans, the fine-grained classification of crop diseases is common, but remains challenging. In general, different sub-categories have very similar appearance, although occasionally the different sub-categories are completely inconsistent. More seriously, the many visual disturbances (such as reflection, dispersion, and blur) caused by dew, shooting jitter, and light intensity seriously reduce the classification accuracy of crop disease images (Lu et al., 2017).

In terms of both theoretical research and practical applications, the fine-grained image classification of crop leaf diseases is of great importance, and is thus the focus of this study. Many researchers have studied the classification of crop diseases based on pattern recognition and machine learning. Guo et al. (2014) utilized texture and color features using a Bayesian approach for recognizing downy mildew, anthracnose, powdery, and gray mold infection with respective accuracy levels of 94.0, 86.7, 88.8, and 84.4%. Zhang et al. (2017) developed a leaf disease identification application in cucumber plants. This application isolates the infected part of the leaf through k-means clustering before extracting the color and shape, resulting in an accuracy level of 85.7%.

Although the above methods have made some progress, the identification and classification of diseases of different crops under actual field conditions can be further improved. For example, although some models can achieve extremely high accuracy on datasets under laboratory conditions, they often have poor identification effects when faced with actual field conditions. We think this is because insufficient disease features are extracted, resulting in a lack of disease details. In summary, the main challenge of fine-grained image classification of crop leaf diseases is undoubtedly the subtle discrimination between different sub-categories. The primary difficulties can be roughly divided into three aspects: (1) the similarity between the sub-categories under the same disease category is very strong; (2) the field environment has significant background interference; and (3) the location of different crop diseases is inconsistent.

In an attempt to overcome these difficulties, many researchers have applied convolutional neural network (CNN) to crop disease classification. To investigate the impact of dataset size and species on the effectiveness of crop disease classification based on deep learning and transfer learning, Barbedo (2018) showed that, although CNNs can largely overcome the technical limitations associated with automated crop disease classification, training with a limited set of image data can have many negative consequences. Kaya et al. (2019) studied and demonstrated that the transfer learning model can help crop classification identification and improve the low-performance classification model. Too et al. (2019) fine-tuned and evaluated the most advanced deep CNN for image-based crop disease classification. The data used in their experiments covered 38 different categories, including disease and health images of the leaves of 14 crops from PlantVillage. The accuracy of DenseNet reached 99.75%, better than that of other models. Cruz et al. (2019) used CNNs to detect leaf images of Grapevine Yellows (GY) disease in red vines (cv. Sangiovese). ResNet-50 was found to be the best compromise network in terms of accuracy and training cost. Turkoglu et al. (2019) proposed a multi-model pre-trained CNN (MLP-CNN) based on long short-term memory for detecting apple diseases and insect pests. Their results were comparable to or better than those of pre-trained CNN models. Deep learning has been widely applied to various crop categories and crop disease classification studies, and deep learning models based on transfer learning can accelerate the training stage. At the same time, to cope with the impact of complex scenes on model classification performance, it is necessary to enhance the performance of CNNs to better handle fine-grained image classification tasks.

In recent years, it has been found that human cognitive processes do not focus attention on the entire scene at one time. On the contrary, they pay more attention to local regions in the scene while extracting relevant information. Models based on attention mechanisms have achieved good results on many challenging tasks, such as visual question answering (Malinowski et al., 2018), object detection (Li et al., 2019), and scene segmentation (Fu et al., 2019). Although the attention mechanism has been applied to different tasks, it has not been used for the fine-grained classification of crop disease images.

In this research, we propose a novel attention mechanism and use transfer learning to quickly build several fine-grained image classification models of crop diseases based on the attention mechanism, so as to solve the problem that the accuracy of CNN model in complex scenes is low due to visual interference in practical applications. Therefore, the contributions of this paper are as follows:

According to the characteristics of crop disease images in real scenes, a fine-grained fine-tuning classification algorithm based on attention mechanism is constructed on the basis of using pre-trained CNN to extract convolutional features of fine-grained images as the input of the network. The attention mechanism makes the classification algorithm pay more attention to some more discriminative local regions of the image, thereby improving the classification accuracy of the model in complex scenes.

We collect crop disease images in real scenes and added these images to the PlantVillage dataset to form a new

hybrid dataset for training the CNN model. We verify the effectiveness of our proposed method by designing multiple comparative experiments.

In addition, we also explore and prove the importance of the image source to the classification results, as well as the impact of problem scenes and special interference on the classification results.

The rest of this paper is organized as follows. Section Materials and methods introduces the main experimental dataset and our proposed fine-grained fine-tuning classification algorithm based on the attention mechanism. The experimental results are described in section Results. In section Discussion, we discuss the importance of training image sources, the impact of problem scenes and special interference on the classification results. Finally, this paper concludes in Section conclusion.

MATERIALS AND METHODS

Image Datasets

This study considered the PlantVillage public dataset of 52,629 images (except for images from the Tomato Two Spotted Spider Mite, Tetranychus urticae, category) (Mohanty et al., 2016), which covers a total of 37 categories including images of 14 different healthy or diseased crops. Using the Scrapy web crawler on the Internet's agricultural technology and consulting platforms, we then extracted a total of 5,571 images uploaded by users in the abovementioned 37 categories (images of crop diseases under actual field conditions). Finally, CNN models were trained and tested using the full dataset of 58,200 healthy and diseased crop disease images (**Figure 1**).

Table 1 provides statistical data for the 37 categories of the dataset, such as the number of images for each category and the percentage of images taken under laboratory or field conditions. It is well-known that there is a single color or no background in the image of laboratory conditions, while the background in the image of field conditions is relatively complex and changeable. As shown in **Table 1**, nearly 10% of the available images were taken under field conditions.

Figure 2 shows disease images of potato late blight, including four disease images obtained under field conditions and four disease images obtained under laboratory conditions. The increase in complexity of the four diseased images under field conditions is obvious (e.g., there are many leaves and other parts in the images, different backgrounds, shadow effects, and so on).

The dataset includes images taken under laboratory conditions and under field conditions (see **Figure 2**); the percentages of each are presented in **Table 1**. The whole dataset was randomly divided into a training set (80%) and a test set (20%). Therefore, 46,560 images were used for CNN model training, while the remaining 11,640 images were used to test the performance of the model. The training set and the test set were preprocessed to satisfy the model's input size requirements, and the image sizes were reduced and cropped to 256×256, 299×299, and 331×331 pixels.

We conducted several experiments to evaluate the importance of the conditions under which the leaf images were captured. Namely, we first conducted the training using only laboratory conditions images (PlantVillage, 52,629 photos) and the testing using images of actual field conditions (Internet, 5,571 photos), and then performed training using only images of actual field conditions (Internet, 5,571 photos) and testing with images taken under laboratory conditions (PlantVillage, 52,629 photos).

Experimental Methods and Parameters

Transfer Learning

VGG, ResNet, and other deep CNN models have achieved great success in image classification. The pre-trained deep CNN model has been fully trained on a large image dataset (ImageNet), allowing many features required for image classification to be learned. Therefore, we can use the idea of transfer learning to fully utilize the large amount of knowledge learned by pre-training the CNN model on the ImageNet dataset, and apply it to crop disease image classification. This paper describes how the transfer learning method of parameter transfer was adopted to remove the maximum pooling and fully connected layers after the final convolution, and introduces a new fine-grained classification model based on the attention mechanism. Compared with the random initialization of the weight parameters of each layer of the network, the fine-tuning method helps accelerate the convergence of the network.

For image classification, there are several CNN baseline models that have been successfully applied to specific tasks. Regarding the task of image recognition and classification of crop diseases, six CNN models that were pre-trained using ImageNet have been applied: (1) VGG16 and VGG19 (Simonyan and Zisserman, 2015), (2) ResNet50 (He et al., 2016), (3) InceptionV3 (Szegedy et al., 2016), (4) Xception (Chollet, 2017), and (5) NASNetLarge (Zoph et al., 2018). The training and testing processes of these pre-trained models and of the proposed fine-grained pre-trained model based on the attention mechanism were implemented using the TensorFlow machine learning computing framework. Model training and testing was conducted with four NVIDIA Tesla V100 GPUs.

Attention Mechanism

The attention mechanism was first applied to natural language processing. It is often combined with recurrent neural networks, resulting in good prediction and processing ability for text sequences. In recent years, the attention mechanism has also been widely used for image classification (Meng and Zhang, 2019; Xiang et al., 2020), object detection (Chen and Li, 2019; Xiao et al., 2020), and image description generation (Liu M. et al., 2020; Zhang et al., 2020). In the field of crop disease classification, most researchers have tended to use transfer learning technology. There has also been some research on crop disease identification based on the attention mechanism (Nie et al., 2019; Karthik et al., 2020). These previous studies have focused on a certain crop, and so the disease category and scale of the dataset are limited. Therefore, we conducted related experiments to verify that increasing the attention mechanism can improve the effect of crop disease classification based on transfer learning technology.

The attention of an image refers to the process of obtaining a target region that requires attention as the human eye rapidly scans the global image. This target region is assigned more

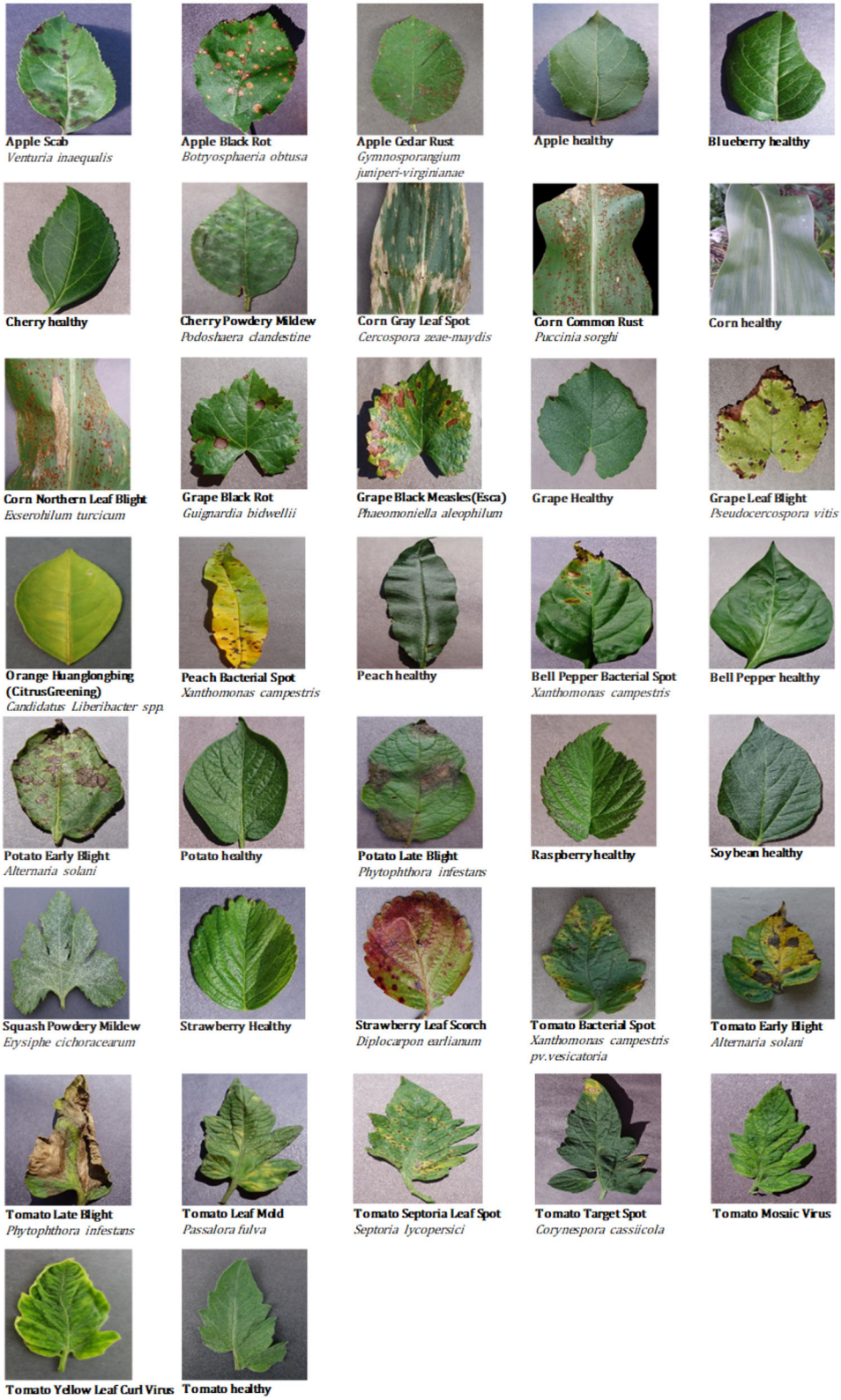

FIGURE 1 | Examples of crop leaf images in the dataset.

TABLE 1 | Statistics of crop healthy/diseased images and related data.

Class	Plant common name	Disease common name	Disease scientific name	Images (PlantVillage)	Laboratory conditions (%)	Field conditions (%)
1	Apple	Apple scab	*Venturia inaequalis*	800 (630)	78.75%	21.25%
2	Apple	Black rot	*Botryosphaeria obtuse*	800 (621)	77.63%	22.38%
3	Apple	Cedar apple rust	*Gymnosporangium juniperi-virginianae*	400 (275)	68.75%	31.25%
4	Apple	—	—	1,800 (1,645)	91.39%	8.61%
5	Blueberry	—	—	1,700 (1,502)	88.35%	11.65%
6	Cherry (and sour)	—	—	1,000 (854)	85.40%	14.60%
7	Cherry (and sour)	Powdery mildew	*Podosphaera* spp.	1,200 (1,052)	87.67%	12.33%
8	Com (maize)	Cercospora leaf spot	*Cercospora zeae-maydis*	700 (513)	73.29%	26.71%
9	Com (maize)	Common rust	*Puccinia sorghi*	1,300 (1,192)	91.69%	8.31%
10	Com (maize)	—	—	1,300 (1,162)	89.38%	10.62%
11	Com (maize)	Northern Leaf Blight	*Exserohilum turcicum*	1,100 (985)	89.55%	10.45%
12	Grape	Black rot	*Guignardia bidwellii*	1,300 (1,180)	90.77%	9.23%
13	Grape	Esca (Black measles)	*Phaeomoniella chlamydospora*	1,500 (1,383)	92.20%	7.80%
14	Grape	—	—	600 (423)	70.50%	29.50%
15	Grape	Leaf blight	*Pseudocercospora vitis*	1,200 (1,076)	89.67%	10.33%
16	Orange	Huanglongbing	Candidatus Liberibacter	5,700 (5,507)	96.61%	3.39%
17	Peach	Bacterial sport	*Xanthomonas campestris*	2,400 (2,297)	95.71%	4.29%
18	Peach	—	—	500 (360)	72.00%	28.00%
19	Pepper, bell	Bacterial spot	*Xanthomonas campestris*	1,100 (997)	90.64%	9.36%
20	Pepper, bell	—	—	1,600 (1,478)	92.38%	7.63%
21	Potato	Early blight	*Altemaria solani*	1,200 (1,000)	83.33%	16.67%
22	Potato	—	—	300 (152)	50.67%	49.33%
23	Potato	Late blight	*Phytophthora infestans*	1,200 (1,000)	83.33%	16.67%
24	Raspberry	—	—	500 (371)	74.20%	25.80%
25	Soybean	—	—	5,200 (5,090)	97.88%	2.12%
26	Squash	Powdery mildew	*Erysiphe cichoracearum*, *Sphaerotheca fuliginea*	2,000 (1,835)	91.75%	8.25%
27	Strawberry	—	—	600 (456)	76.00%	24.00%
28	Strawberry	Leaf scorch	*Diplocarpon earlianum*	1,300 (1,109)	85.31%	14.69%
29	Tomato	Bacterial spot	*Xanthomonas campestris* pv. Vesicatoria	2,300 (2,127)	92.48%	7.52%
30	Tomato	Early blight	*Altemaria solani*	1,200 (1,000)	83.33%	16.67%
31	Tomato	Late blight	*Phytophthora infestans*	2,100 (1,909)	90.90%	9.10%
32	Tomato	Leaf Mold	*Fulvia fulva*	1,100 (952)	86.55%	13.45%
33	Tomato	Septoria leaf spot	*Septoria lycopersici*	1,900 (1,771)	93.21%	6.79%
34	Tomato	Target spot	*Corynespora cassiicola*	1,600 (1,404)	87.75%	12.25%
35	Tomato	Tomato mosaic virus	*Tomato mosaic virus* (ToMV)	500 (373)	74.60%	25.40%
36	Tomato	TYLCV	Begomovirus (Fam. Geminiviridae)	5,500 (5,357)	97.40%	2.60%
37	Tomato	—	—	1,700 (1,591)	93.59%	6.41%
TOTAL:				58,200 (52,629)	90.43%	9.57%

Images column gives the number of images in each category, where the figures in parentheses are from the public dataset.

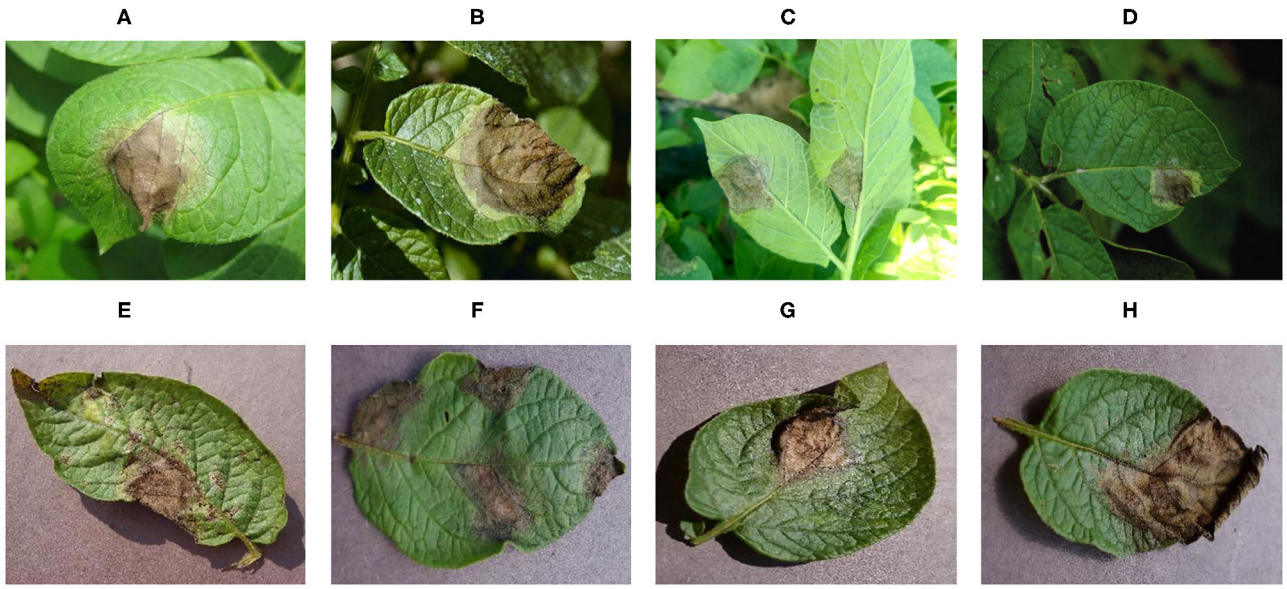

FIGURE 2 | Image sample of potato late blight. **(A–D)** Field conditions, **(E–H)** laboratory conditions.

attention (weight distribution) to obtain the required detailed information about the target, with other useless information suppressed. Soft attention is most commonly used, because this is a completely differentiable process that can realize end-to-end learning in CNN models. Most soft attention models learn an attention template to align the weights of different regions in a sequence or an image and use this template to locate the distinguishable regions. Different from soft attention, the hard attention mechanism is a random, non-differentiable process that determines the importance of individual regions one at a time, rather than identifying the important regions within the whole image.

For image classification, the weight of the arithmetic mean of attention can be extracted through attention learning to form the attention spectrum of the image. Similar to traditional natural language processing, the image-based attention can be obtained through the model illustrated in **Figure 3**.

In **Figure 3**, I is the input image. The attention model has n parameters, $a_1, a_2, \cdots, a_i, a_n$, which respectively represent a description of each part of the image. O is the return value of the model's attention spectrum (more specifically, the weight values of the n parameters), which is determined from the importance of each a_i relative to the input I. By filtering the input image through this output, the region that requires most attention can be identified.

Proposed Model

Based on the pre-trained model described in Section transfer learning and the attention mechanism introduced in Section attention mechanism, this paper proposes a fine-grained classification model based on the attention mechanism (**Figure 4**). By learning the attention of the CNN feature spectrum, the attention model calculates and identifies the most important region of the feature spectrum for the final classification task, and provides the maximum attention input (weight distribution). However, adding the attention weight to the last layer of the CNN features will cause different degrees of suppression of the original features. To overcome this suppression, the weighted feature spectrum is added to the original feature spectrum. The fusion spectrum is then input into the fully connected layer. In the second fully connected layer, the attention feature spectrum transformed by the global average pooling dimension is connected with the fully connected feature spectrum in the channel direction, before being sent to the classification layer for classification.

The attention model proposed in this paper adopts an unsupervised training mode. There is no pre-labeled ground truth to constrain the attention spectrum, so there is no separate loss calculation. Instead, a backpropagation adaptive mode is used to constrain the weight distribution of attention. The loss function defined in this paper is expressed as:

$$loss = \frac{1}{2n} \sum ||y_{truth}(x) - y_{pred}(x)||^2 \quad (1)$$

where $n = 37$ is the number of input samples, x is the input sample, y_{truth} is the actual category, and y_{pred} is the predicted category output by the final layer of the network. In the process of backpropagation, the output error of the *Softmax* layer is backpropagated, and the parameters are updated using the random gradient descent method, so that the final loss function value decreases and the network converges.

Through the soft attention mechanism, the output of the final convolutional layer of the CNN is obtained. This is taken as the input of the attention model, and the corresponding attention

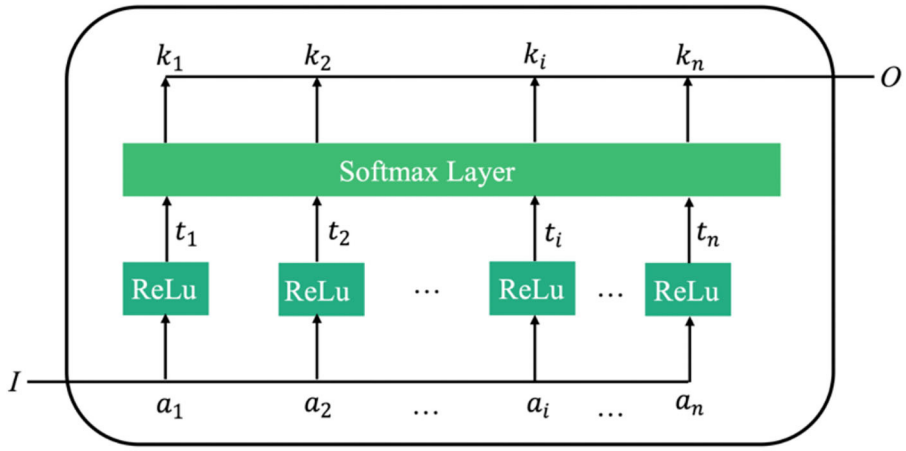

FIGURE 3 | Schematic diagram of learning mechanism of attention.

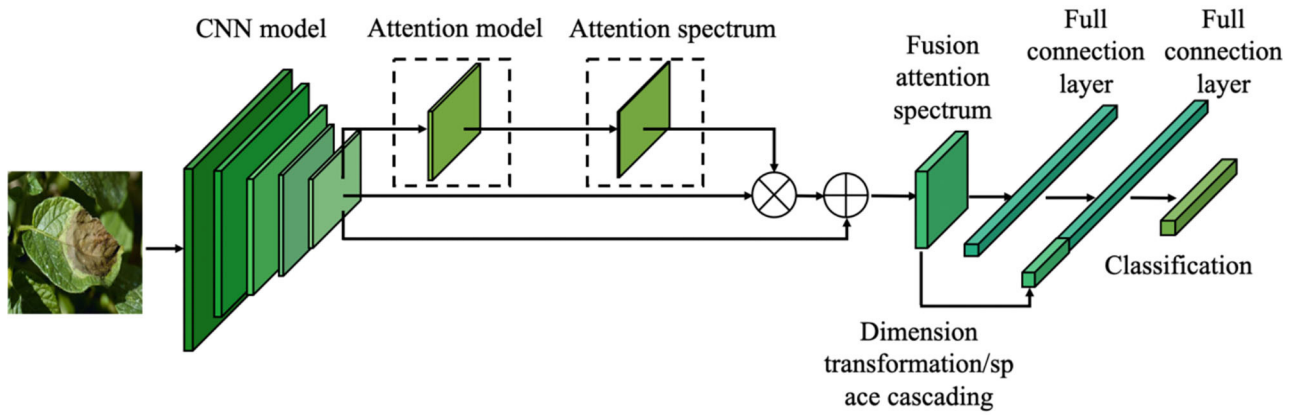

FIGURE 4 | Fine-grained image classification network structure of crop diseases based on attention mechanism.

spectrum is calculated. The original feature spectrum is then weighted by the attention spectrum, and the output attention feature spectrum is provided as the input for the subsequent network. Let the feature spectrum of the output of the final convolution spectrum after the pooling operation be expressed as $f \in \mathbb{R}^{H\times W\times C}$, where H and W refer to the height and width of the feature spectrum of this layer, C refers to the number of channels of the feature spectrum of this layer, and for each position (m, n) on the spectrum, its feature value is expressed as $f_{m,n}\epsilon\mathbb{R}^C$. The corresponding attention weight $W_{m,n}$ can then be obtained as:

$$W_{m,n} = ATT(f_{m,n};\ W_{att}) \tag{2}$$

where ATT is a mapping function learned by the attention model and W_{att} is the weight parameter of the attention model. Through $Softmax$ regression of $w_{m,n}$, the final attention spectrum $M = [M_{m,n}]$ is obtained as a normalized probability matrix, where $M_{m,n}$ is expressed as:

$$M_{m,n} = Softmax(W_{m,n}) \tag{3}$$

As can be seen from **Figure 5**, the attention model proposed in this paper takes the output of the final convolution spectrum in the neural network as its input. The attention model includes two convolutional layers and one *Softmax* layer. The kernel sizes of the convolutional layers are 3×3 and 1×1. The attention feature spectrum $f^{att} = [f^{att}_{m,n}]$ is obtained by multiplying the attention spectrum M by the CNN feature spectrum f, and is expressed as

$$f^{att} = \left[f^{att}_{m,n}\right] = [M_{m,n} \cdot f_{m,n}] \tag{4}$$

The 3×3 convolution kernel further extracts the CNN feature of the final convolution spectrum. In order not to reduce the feature receptive field and feature information, a convolution kernel with the same size as the original network is selected. Compared with the 3×3 convolution kernel, the 1×1 convolution kernel enables information interaction and integration across channels. By connecting features in the channel direction, nonlinear components can be added to features to improve the feature expression ability of the attention model.

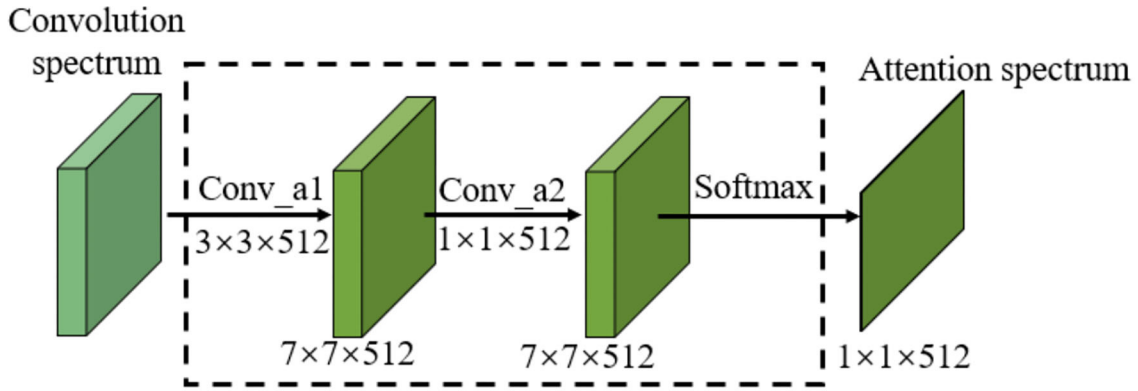

FIGURE 5 | Attention calculation model for images.

The attention spectrum of the final convolution spectrum of the CNN is obtained through the attention model, and the attention spectrum and the original CNN feature spectrum are then multiplied to obtain the attention feature spectrum. The attention spectrum is the spectrum obtained after normalizing the weights of the features. According to this definition, the attention feature spectrum obtained after multiplication has a certain attenuation compared with the original CNN feature spectrum. Additionally, during the convolution and probability calculation, the spatial transformation of the CNN feature spectrum and noise addition means that the calculated attention spectrum may be distorted. In this case, the obtained attention spectrum has no guiding significance for the original image spectrum. To overcome this problem, once the attention feature spectrum has been obtained, the original CNN feature spectrum is added and fused to obtain the final attention f_{all}^{att}, which is input into the subsequent fully connected layer, as shown in Equation (5).

$$f_{all}^{att} = f + M \cdot f \tag{5}$$

By adding the attention spectrum to the CNN feature spectrum, the distortion of the attention spectrum is overcome and the original feature spectrum before the fully connected layer can be effectively utilized.

By extracting and merging the attention spectrum, a spectrum of features is obtained that is well-located and noticeable in space. This spectrum is then input to the subsequent fully connected layer. As the connection operation maps the convolutional spectra of all channels to one point in the fully connected layer, the spatial information is destroyed by the operation of the fully connected layer. The original intention of introducing the attention model is to extract and improve the significant regions of the CNN feature spectrum in space. However, after the fully connected layer, the spatial information of the extracted attention feature has also been destroyed. Therefore, the attention space feature is reused by connecting the attention feature spectrum in the final fully connected layer, as shown in **Figure 6**.

Evaluation of the Model

The accuracy, precision (P), recall (R), and comprehensive F_1 evaluation index were used to evaluate the crop disease image classification model. The F_1 value is the harmonic average of the precision and recall, and has a maximum of 1 and a minimum of 0. It is calculated as follows:

$$F_1 = \frac{2PR}{P+R} \times 100\% \tag{6}$$

Experimental Details

In all our experiments, we preprocessed the images to sizes of 256×256, 299×299, and 331×331 pixels, conducted a total of 1,000 training epochs, and used a batch size of 32. We used a momentum SGD initial learning rate of 0.001. When the standard evaluation stopped increasing, the learning rate was multiplied by 0.1 until it had dropped to 0.0001. After lowering the learning rate, we waited for five epochs before returning to normal operation. If the loss of the test set did not improve after 20 epochs, the learning rate was reduced. We conducted experiments using multiple pre-trained models, all of which are robust to the selection of hyper-parameters.

RESULTS

Compared With the Pre-training Model and the Effect of the Attention Mechanism

Tables 2, **3** present the classification accuracy, precision, recall, and F_1 values of various models on the test set. The results indicate that the fine-grained fine-tuning classification models based on the attention mechanism outperform the original pre-trained models by 1–2% in terms of accuracy, precision, recall, and F_1 value. This demonstrates that the attention mechanism improves the classification performance of the models and allows them to focus on key regions in the image. The fine-grained NASNetLarge model based on the fine-grained attention mechanism achieves the highest accuracy, precision, recall, and F_1 values, and thus provides the best classification performance. These 12 models were further trained using only the original image to record the training period for the best performance. As

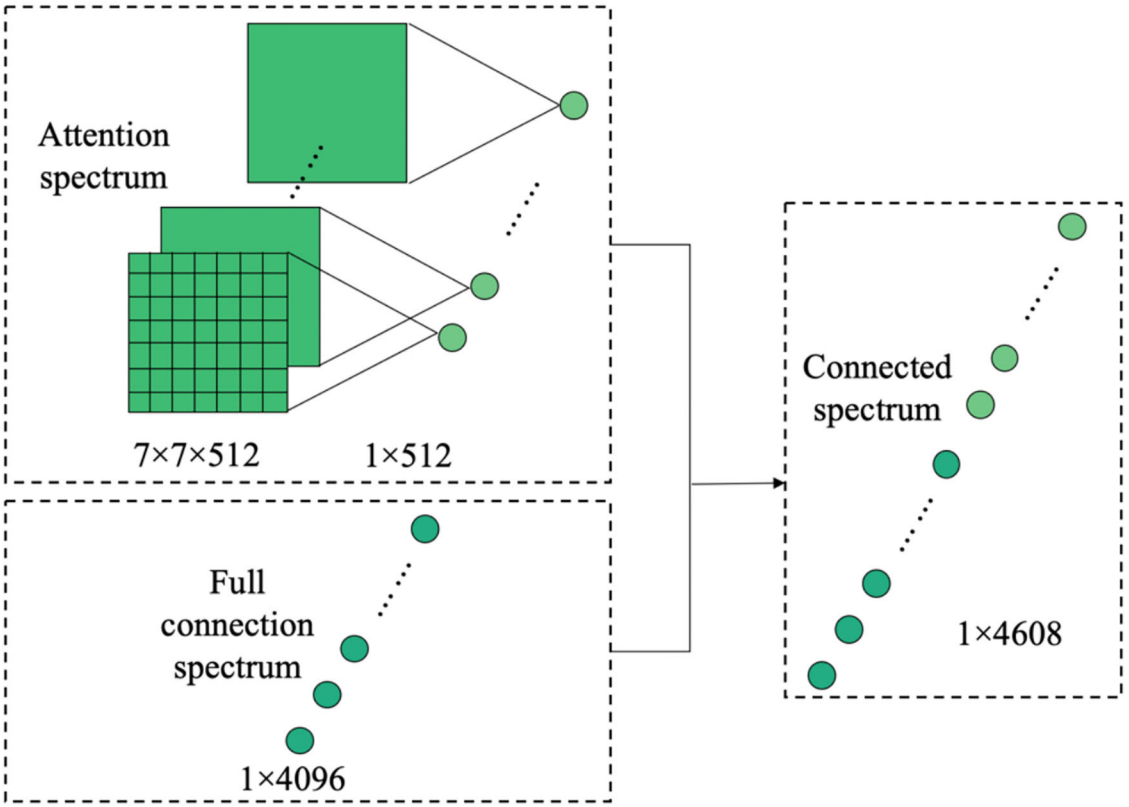

FIGURE 6 | Dimension transformation and connection diagram of attention feature spectrum.

shown in **Table 3**, the fine-grained NASNetLarge model based on the attention mechanism achieves the highest classification accuracy of 95.62%. Thus, this model was used in subsequent experiments for crop disease image classification.

Figure 7 provides a visual representation of some random images from the test set. The table on the left of the original image shows the predicted classification. The image to the right of the original image is a visual representation of the attention mechanism using the fine-grained fine-tuning NASNetLarge model based on the attention mechanism. The highest-ranked classification result for each image was considered as the final classification result predicted by the model. The images of the crop leaves shown in **Figure 7** are correctly classified. In most cases, the degree of certainty for the correct classification is close to 100%, so there is no actual ranking.

Testing on Different Dataset

We also comprehensively evaluated our algorithm on public plant datasets of Flavia (Wu et al., 2007), Swedish Leaf (Söderkvist, 2001), and UCI Leaf (Silva et al., 2013). These datasets contain clear images, and they are widely used datasets in this field, often used for algorithm development and comparison. The statistics of three datasets are shown in **Table 4**. We follow the same training/test split as in Section image datasets.

The Flavia dataset contains 1,907 images of 32 species of plants. All images in the dataset have a white background, and the number of each category varies from dozens of images and is relatively unbalanced.

The Swedish Leaf dataset contains 15 plant species, with 75 images in each category. All plant leaf images are images with white background, and the quality and resolution of each image is high.

The UCI Leaf dataset contains 40 different plants and a total of 443 images. The background colors of the images in this dataset are all pink. The number of images in each category ranges from a few to a dozen.

As seen from **Table 5** the NASNetLarge model based on the attention mechanism constructed by our proposed method can still get the best classification accuracy on the three public plant datasets. Therefore, it can be proved that our model has better performance across datasets and can achieve efficient classification on datasets of different sizes.

Validation and Comparison of Proposed CNN With Traditional Machine Learning Models

The traditional machine learning methods used for comparison in this paper are SVM, Decision tree, k-NN, and Naive Bayes. Features such as Hu-moments, Haralick features, LBP features,

TABLE 2 | Results of pre-trained model on test set.

Pre-trained model	Size(MB)	Accuracy (%)	Precision (%)	Recall (%)	F1-measure (%)	Parameters
VGG16 (Simonyan and Zisserman, 2015)	56.16	83.07	82.52	80.13	81.31	15,360,589
VGG19 (Simonyan and Zisserman, 2015)	76.42	85.79	83.45	81.75	82.59	20,670,285
ResNet50 (He et al., 2016)	90.38	85.82	85.21	83.63	84.41	25,769,613
InceptionV3 (Szegedy et al., 2016)	83.84	88.47	87.78	85.47	86.61	23,984,685
Xception (Chollet, 2017)	79.81	91.22	90.24	87.05	88.62	23,043,381
NASNetLarge (Zoph et al., 2018)	327.69	92.78	92.16	90.83	91.49	89,082,719

TABLE 3 | Results of fine-grained classification model based on attention on test set.

Model based on attention mechanism	Size(MB)	Accuracy (%)	Precision (%)	Recall (%)	F1-measure (%)	Parameters
VGG16*	59.23	85.53	84.68	81.32	82.97	15,514,783
VGG19*	79.75	86.40	84.93	82.05	83.47	20,823,653
ResNet50*	93.91	87.03	86.27	85.32	85.79	25,942,746
InceptionV3*	87.53	90.64	88.51	88.48	88.49	24,173,257
Xception*	83.28	92.89	91.82	90.95	91.38	23,218,425
NASNetLarge*	331.75	95.62	94.35	91.79	93.05	89,286,814

*The * indicates that the pre-training model is used.*

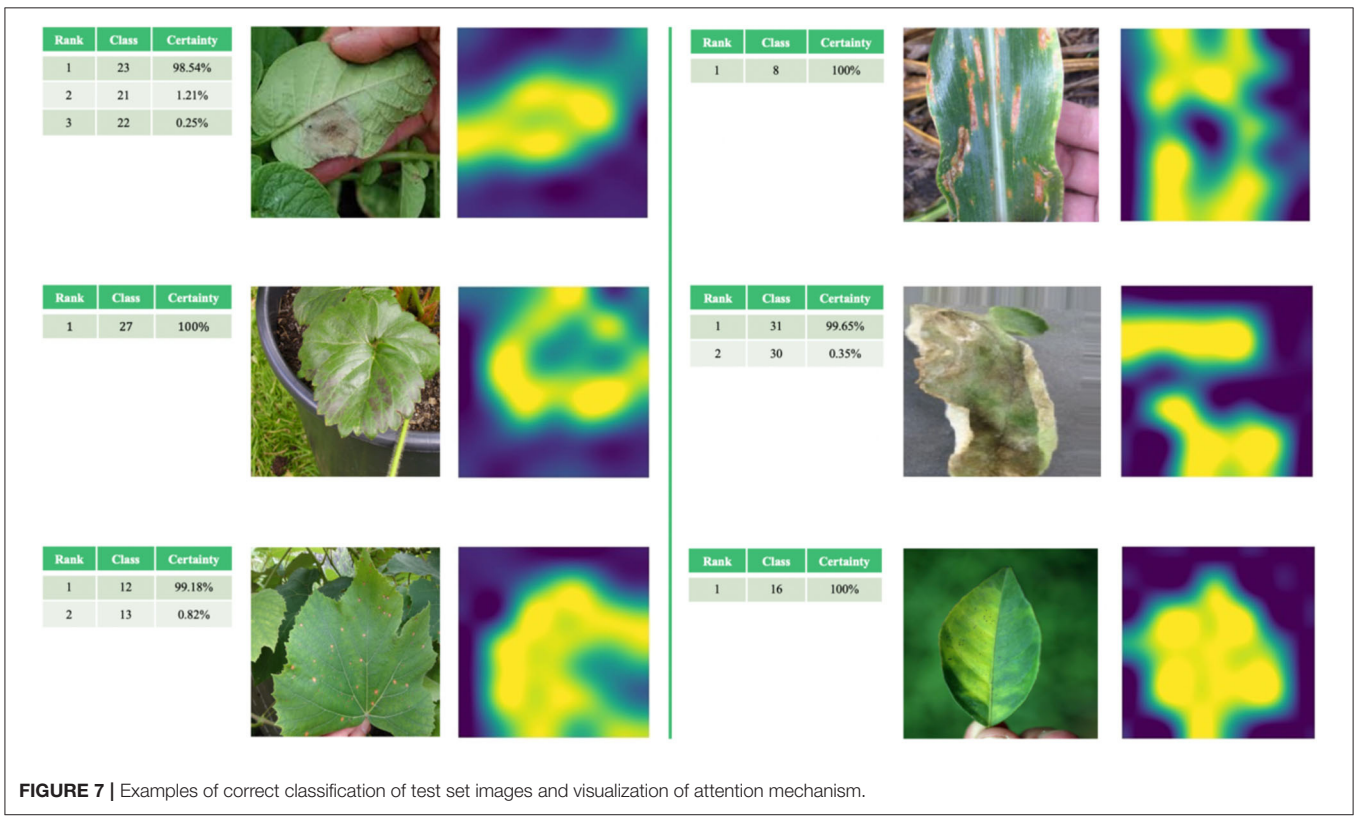

FIGURE 7 | Examples of correct classification of test set images and visualization of attention mechanism.

and HSV features have been used to evaluate the performance of all traditional Machine Learning algorithms. The results are given in **Table 6**.

Table 6 shows that the accuracy, precision, recall and F_1 of our proposed model are much higher than those obtained using other machine learning algorithms.

TABLE 4 | Statistics of benchmark datasets.

Datasets	Class	Train	Test
Flavia	32	1,526	381
Swedish leaf	15	900	225
UCI leaf	40	356	87

TABLE 5 | Comparison of methods for image classification in three datasets.

References	Datasets	Accuracy/%	Method
Lee et al. (2017)	Flavia	99.40	CNN, Fine-tuning
Yousefi et al. (2017)	Flavia	97.50	Fourier and Wavelet Descriptors, MLP
Murat et al. (2017)	Flavia	95.25	HOG, Moments, ANN,
	Swedish	99.89	RF, and SVM
Kaya et al. (2019)	Flavia	99.00	DF – VGG16/LDA
	Swedish	98.80	CNN – RNN
	UCI Leaf	96.20	DF – Alexnet/LDA
Our	Flavia	99.72	NASNetLarge –
	Swedish	99.90	Attention
	UCI Leaf	98.74	

DISCUSSION

Importance of Training Image Type

The fine-grained NASNetLarge model based on the attention mechanism produced the best classification effect, and was therefore further tested to study the importance of assessing the conditions under which the leaf image was captured. The corresponding results are presented in **Table 7**.

When only laboratory or field condition images are used for training, the accuracy on the test set is significantly lower than when both laboratory and field condition images are used for training. The results show that the model can obtain better performance when using images obtained under field conditions for training and requiring classification of laboratory condition images (F_1 value is nearly 60.47%). In contrast, when the laboratory condition images are used for training and the field condition images are classified, the classification performance is obviously reduced (F_1 value is about 34.11%). This shows that image classification under field conditions is a more difficult and complicated task than the classification of images obtained under laboratory conditions, and proves that the construction of an efficient automatic detection and diagnosis system for crop diseases using images obtained under field conditions is of great significance.

Problematic Situations and Indicative Cases

The fine-grained NASNetLarge model based on the attention mechanism reached an accuracy level of 95.62% on the test set of 11,640 images, of which 11,130 images were correctly classified. Among the 4.38% of misclassified images, there are some "problematic" images that do not contain crop leaves at all (as shown in **Figures 8A,B**). These images should be classified into category 31 (Tomato late blight, phytophthora infestans), but the model classifies their predictions into category 10 (Corn healthy), as shown in the classification table in **Figure 8**. The classification table shows the predicted classification ranking output of the final model on the original image. These images are misclassified by the model (the "correct" classification would be category 31). In fact, they do not belong to any category, because there are no crop leaves in the image. However, they are all classified as category 10. We infer that the images in category 10 (**Figures 8C,D**) contain similar soils, while the corn leaves are very slender and occupy a small portion of the image. If such problematic examples are excluded, the accuracy of the final model will be higher than 95.62%.

There are several other problems with the images obtained under field conditions, including: (1) shadows on the leaves in the images, with some images appearing dark and shaky; (2) other objects in the image that are not related to the leaf itself, such as a trunk, fruit, or fence. Note that these problematic images occupy a very small portion of the dataset. In short, according to the certainty levels provided by the final model, the attention mechanism-based approach proposed in this paper overcomes these problems in most cases.

A typical case is category 8 (Corn cercospora leaf spot, cercospora zeae-maydis). **Figure 9** shows the classification results of the model for eight representative images of category 8, including four incorrectly classified images (the lower four images in **Figure 9**) and four correctly classified images (the upper four images in **Figure 9**). The first three upper images were correctly classified, with a certainty of ∼100%, while the fourth image was correctly classified with a certainty of ∼79% (the second ranking, with a certainty level of 16% for category 10, corresponds to corn crops with different diseases). The four misclassified images (the lower four images in **Figure 9**) include a wide range of partial shadows or complex backgrounds, which increase the misclassification rate of the model. For the middle two lower images, the correct classification is ranked second, while the first ranking is category 10 or 11 (corn crop diseases). Therefore, the model correctly identifies the crop species, but does not accurately detect the particular crop disease.

CONCLUSIONS

This study constructed, trained, and tested a fine-grained neural network model based on the attention mechanism for the classification of simple leaves of healthy or diseased crops. The model was trained using 58,200 publicly available images obtained under both laboratory conditions and field conditions. The data include 14 crop species in 37 different categories of [crop, disease] combinations, including some healthy crops. The optimal model was found to be a fine-grained NASNetLarge neural network based on the attention mechanism, which achieved an accuracy level of 95.62% (precision 94.35%, recall 91.79%, F_1 value 93.05%) in the classification of the 11,640 images

TABLE 6 | Comparison of methods for image classification in three datasets.

Models	Features	Accuracy/%	Precision/%	Recall/%	F1-measure/%
Naïve Bayes	Haralick, Hu	31.67	17.11	12.23	14.26
	Hu-moments, HSV	35.61	20.92	15.56	17.84
	Haralick, Hu, HSV	39.08	29.33	20.55	24.17
	Haralick, Hu, HSV, LBP	43.24	36.76	25.64	30.21
Decision tree	Haralick, Hu	41.33	23.24	18.85	20.82
	Hu-moments, HSV	47.18	30.52	25.21	27.61
	Haralick, Hu, HSV	53.91	43.45	34.69	38.58
	Haralick, Hu, HSV, LBP	60.23	48.32	41.36	44.57
SVM	Haralick, Hu	43.33	30.42	23.28	26.38
	Hu-moments, HSV	52.67	41.23	33.01	36.66
	Haralick, Hu, HSV	55.82	45.35	36.82	40.64
	Haralick, Hu, HSV, LBP	61.45	54.51	43.24	48.23
k-NN	Haralick, Hu	69.66	65.23	53.72	58.92
	Hu-moments, HSV	75.38	73.57	61.34	66.90
	Haralick, Hu, HSV	79.52	77.41	66.38	71.47
	Haralick, Hu, HSV, LBP	84.45	80.65	74.77	77.60
Our model (NASNetLarge*)		95.62	94.35	91.79	93.05

*The * indicates that the pre-training model is used.*

TABLE 7 | Results of using different training sets and test sets with the optimal model under laboratory conditions and field conditions.

Model based on attention mechanism	Training: laboratory Testing: actual field conditions				Training: actual field conditions Testing: laboratory			
	Accuracy (%)	Precision (%)	Recall (%)	F1-measure (%)	Accuracy (%)	Precision (%)	Recall (%)	F1-measure (%)
NASNetLarge*	38.52	34.28	33.95	34.11	66.85	61.32	59.64	60.47

*The * indicates that the pre-training model is used.*

in the test set. The fine-grained NASNetLarge neural network model based on the attention mechanism achieves excellent classification performance by analyzing simple leaf images, so it is highly suitable for the automatic detection and diagnosis of crop diseases. In addition, the experimental results show that the images taken under field conditions in the training set are of high importance, indicating that when training such models, the proportion of images obtained under field conditions in the training set should be carefully considered.

For the backbone network of NASNetLarge, the results show that the NASNetMobile neural network model, similar to NASNetLarge, achieves state-of-the-art classification results on related datasets, surpassing the performance of previous lightweight networks such as MobileNet (Sandler et al., 2018) and ShuffleNet (Ma et al., 2018). As NASNetMobile requires little computing power to classify the given images, it can run on mobile devices such as smartphones, drones, or automatic agricultural vehicles for real-time monitoring and disease identification of large open-air crops. At present, due to the large-scale application of 5G, high-efficiency transmission, and improvements to the hardware configuration of mobile terminal equipment, it is possible to upload images locally to the cloud server for processing, and then return the identification and classification results to the terminal (Johannes et al., 2017; Toseef and Khan, 2018; Picon et al., 2019), or to use a GPU/CPU at the terminal to process and display the results (Barman et al., 2020). For growers in remote areas, real-time detection and diagnosis can be carried out through mobile terminals, thus solving the practical problems of obtaining technical crop disease diagnosis and finding experts in the production process. For agricultural technicians, this is equivalent to having a valuable auxiliary consultation tool. In the future, an intelligent crop disease prevention and control recommendation system will be developed based on the results of real-time diagnosis, allowing growers to select different prevention and control methods (e.g., physical or chemical methods) according to the specific conditions. The process and dosage of the methods will also be described in detail. Intelligent crop disease identification and diagnosis, as well as intelligent crop disease prevention and recommendation, will greatly improve production efficiency, realize agricultural, scientific, and technological progress, and push agriculture into the intelligent era.

Although the system developed in this study achieved a high success rate, it is far from becoming a universal tool under actual field conditions (Boulent et al., 2019). At present, the existing research has only considered dozens of [crop, disease] combinations (Ferentinos, 2018), so it is vital to expand the existing database to include more crop species

A

Rank	Class	Certainty
1	10	72.06%
2	11	8.21%
3	8	6.47%
4	26	5.92%
5	32	3.69%
6	31	2.63%

B

C

Rank	Class	Certainty
1	10	86.31%
2	11	3.04%
3	8	2.95%
4	30	1.36%
5	31	~0%

D

FIGURE 8 | (A,B): Images in category 31 and their corresponding "classification" results. **(C,D)**: representative images in category 10.

FIGURE 9 | Examples of correct (upper four images) and incorrect (lower four images) classification of category eight images in the test set (*corn cercospora leaf spot*).

and corresponding diseases. The test set used to evaluate the model was part of the dataset from which the training set was extracted, which is a potential source of bias. This is a common method for training and testing machine learning models. However, to develop a system that can be used effectively in field scenes, data from various sources should

be used for testing to ensure that future users can obtain effective classification results in different scenes (Barbedo, 2018). At present, some preliminary experiments carried out with limited data show that when testing images different from those used in training, the classification performance of the model is significantly reduced to the range of 25–35%. The experimental results show that the classification effect depends on the data source. To improve this, more extensive image datasets should be collected from different geographical areas, field conditions, image capture modes, and multiple sources. Improving the model by increasing the size of the dataset would allow more effective and widespread identification of crop categories and diseases under field conditions.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

GY and YH: conceptualization. GY: methodology, software, formal analysis, data curation, writing—original draft preparation, and visualization. GY and BX: validation. GY and YY: investigation. YH and YY: resources, supervision, and project administration. BX: writing—review and editing. All authors contributed to the article and approved the submitted version.

REFERENCES

Barbedo, J. G. A. (2018). Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. *Comput. Electron. Agric.* 153, 46–53. doi: 10.1016/j.compag.2018.08.013

Barman, U., Choudhury, R. D., Sahu, D., and Barman, G. G. (2020). Comparison of convolution neural networks for smartphone image based real time classification of citrus leaf disease. *Comput. Electron. Agric.* 177:105661. doi: 10.1016/j.compag.2020.105661

Boulent, J., Foucher, S., Théau, J., and St-Charles, P.-L. (2019). Convolutional neural networks for the automatic identification of plant diseases. *Front. Plant Sci.* 10:941. doi: 10.3389/fpls.2019.00941

Chen, H., and Li, Y. (2019). Three-stream attention-aware network for RGB-D salient object detection. *IEEE Trans. Image Process.* 28, 2825–2835. doi: 10.1109/TIP.2019.2891104

Chen, Y., Bai, Y., Zhang, W., and Mei, T. (2019). "Destruction and construction learning for fine-grained image recognition," in *Computer Vision and Pattern Recognition (CVPR)* (Long Beach, CA: IEEE), 5152–5161. doi: 10.1109/CVPR.2019.00530

Chollet, F. (2017). "Xception: deep learning with depthwise separable convolutions," in *Computer Vision and Pattern Recognition (CVPR)* (Honolulu, HI: IEEE), 1800–1807. doi: 10.1109/CVPR.2017.195

Cruz, A., Ampatzidis, Y., Pierro, R., Materazzi, A., Panattoni, A., Bellis, L. D., et al. (2019). Detection of grapevine yellows symptoms in *Vitis vinifera* L. with artificial intelligence. *Comput. Electron. Agric.* 157, 63–76. doi: 10.1016/j.compag.2018.12.028

Ding, Y., Zhou, Y., Zhu, Y., Ye, Q., and Jiao, J. (2019). "Selective sparse sampling for fine-grained image recognition," *Computer Vision and Pattern Recognition (CVPR)* (Long Beach, CA: IEEE), 6599-6608. doi: 10.1109/ICCV.2019.00670

Dubey, A., Gupta, O., Guo, P., Raskar, R., Farrell, R., and Naik, N. (2018). "Pairwise confusion for fine-grained visual classification," in *European Conference on Computer Vision (ECCV)* Lecture Notes in Computer Science, Vol. 11216 (Munich: Springer), 71–88. doi: 10.1007/978-3-030-01258-8_5

Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. *Comput. Electron. Agric.* 145, 311–318. doi: 10.1016/j.compag.2018.01.009

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., et al. (2019). Dual attention network for scene segmentation," in *Computer Vision and Pattern Recognition (CVPR)* (Long Beach, CA: IEEE), 3146–3154. doi: 10.1109/CVPR.2019.00326

Fuentes, A. F., Yoon, S., Lee, J., and Park, D. S. (2018). High-performance deep neural network-based tomato plant diseases and pests diagnosis system with refinement filter bank. *Front. Plant Sci.* 9:1162. doi: 10.3389/fpls.2018.01162

Guo, P., Liu, T., and Li, N. (2014). Design of automatic recognition of cucumber disease image. *Inform. Technol. J.* 13, 2129–2136. doi: 10.3923/itj.2014.2129.2136

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV: IEEE), 770–778. doi: 10.1109/CVPR.2016.90

Johannes, A., Picon, A., Alvarez-Gila, A., Echazarra, J., Rodriguez-Vaamonde, S., Navajas, A. D., et al. (2017). Automatic plant disease diagnosis using mobile capture devices, applied on a wheat use case. *Comput. Electron. Agric.* 138, 200–209. doi: 10.1016/j.compag.2017.04.013

Karthik, R., Hariharan, M., Anand, S., Mathikshara, P., Johnson, A., and Menaka, R. (2020). Attention embedded residual CNN for disease detection in tomato leaves. *Appl. Soft Comput.* 86:105933. doi: 10.1016/j.asoc.2019.105933

Kaya, A., Keceli, A. S., Catal, C., Yalic, H. Y., Temucin, H., and Tekinerdogan, B. (2019). Analysis of transfer learning for deep neural network based plant classification models. *Comput. Electron. Agric.* 158, 20–29. doi: 10.1016/j.compag.2019.01.041

Lee, S. H., Chan, C. S., Mayo, S. J., and Remagnino, P. (2017). How deep learning extracts and learns leaf features for plant classification. *Pattern Recogn.* 71, 1–13. doi: 10.1016/j.patcog.2017.05.015

Li, L., Xu, M., Wang, X., Jiang, L., and Liu, H. (2019). "Attention based glaucoma detection: a large-scale database and CNN model," *Computer Vision and Pattern Recognition (CVPR)* (Long Beach, CA: IEEE), 10571–10580. doi: 10.1109/CVPR.2019.01082

Liu, B., Ding, Z., Tian, L., He, D., Li, S., and Wang, H. (2020). Grape leaf disease identification using improved deep convolutional neural networks. *Front. Plant Sci.* 11:1082. doi: 10.3389/fpls.2020.01082

Liu, M., Li, L., Hu, H., Guan, W., and Tian, J. (2020). Image caption generation with dual attention mechanism. *Inform. Process. Manage.* 57:102178. doi: 10.1016/j.ipm.2019.102178

Lu, Y., Yi, S., Zeng, N., Liu, Y., and Zhang, Y. (2017). Identification of rice diseases using deep convolutional neural networks. *Neurocomputing* 267, 378–384. doi: 10.1016/j.neucom.2017.06.023

Ma, N., Zhang, X., Zheng, H.-T., and Sun, J. (2018). ShuffleNet V2: practical guidelines for efficient CNN architecture design," in *Proceedings of the 15th European Conference* (Munich, Germany), 122–138. doi: 10.1007/978-3-030-01264-9_8

Malinowski, M., Doersch, C., Santoro, A., and Battaglia, P. W. (2018). "Learning visual question answering by bootstrapping hard attention," in *European Conference on Computer Vision (ECCV)*, Lecture Notes in Computer Science, Vol. 11210 (Munich: Springer), 3–20. doi: 10.1007/978-3-030-01231-1_1

Meng, Q., and Zhang, W. (2019). "Multi-label image classification with attention mechanism and graph convolutional networks," in *Proceedings of the ACM Multimedia Asia* (Beijing, China). doi: 10.1145/3338533.3366589

Mohanty, S. P., Hughes, D. P., and Salathe, M. (2016). Using deep learning for image-based plant disease detection. *Front. Plant. Sci.* 7, 1419–1419. doi: 10.3389/fpls.2016.01419

Murat, M., Chang, S.-W., Abu, A., Yap, H. J., and Yong, K.-T. (2017). Automated classification of tropical shrub species: a hybrid of leaf shape and machine learning approach. *Peer J.* 5:e3792. doi: 10.7717/peerj.3792

Mutka, A. M., and Bart, R. S. (2015). Image-based phenotyping of plant disease symptoms. *Front. Plant Sci.* 5:734. doi: 10.3389/fpls.2014.00734

Nie, X., Wang, L., Ding, H., and Xu, M. (2019). Strawberry verticillium wilt detection network based on multi-task learning and attention. *IEEE Access* 7, 170003–170011. doi: 10.1109/ACCESS.2019.2954845

Ou, X., Wei, Z., Ling, H., Liu, S., and Cao, X. (2016). "Deep multi-context network for fine-grained visual recognition," in *International Conference on Multimedia and ExpoWorkshops (ICMEW)* (Seattle, WA: IEEE), 1–4. doi: 10.1109/ICMEW.2016.7574666

Peng, Y., He, X., and Zhao, J. (2018). Object-part attention model for fine-grained image classification. *IEEE Trans. Image Process.* 27, 1487–1500. doi: 10.1109/TIP.2017.2774041

Picon, A., Alvarez-Gila, A., Seitz, M., Ortiz-Barredo, A., Echazarra, J., and Johannes, A. (2019). Deep convolutional neural networks for mobile capture device-based crop disease classification in the wild. *Comput. Electron. Agric.* 161, 280–290. doi: 10.1016/j.compag.2018.04.002

Ramcharan, A., Baranowski, K., McCloskey, P., Ahmed, B., Legg, J., and Hughes, D. P. (2017). Deep learning for image-based cassava disease detection. *Front. Plant Sci.* 8:1852. doi: 10.3389/fpls.2017.01852

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). "MobileNetV2: inverted residuals and linear bottlenecks," in *Computer Vision and Pattern Recognition (CVPR)* (Salt Lake City, UT: IEEE), 4510–4520. doi: 10.1109/CVPR.2018.00474

Silva, P. F. B., Marçal, A. R. S., and da Silva, R. M. A. (2013). "Evaluation of features for leaf discrimination," *in International Conference Image Analysis and Recognition*, Lecture Notes in Computer Science, Vol. 7950 (Póvoa de Varzim: Springer), 197–204. doi: 10.1007/978-3-642-39094-4_23

Simonyan, K., and Zisserman, A. (2015). "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)* (San Diego, CA).

Söderkvist, O. (2001). Computer vision classification of leaves from swedish trees (Master's Thesis). Linkoping University, Linkoping, Sweden, p. 74.

Sun, G., Cholakkal, H., Khan, S., Khan, F., and Shao, L. (2020). "Fine-grained recognition: accounting for subtle differences between similar classes," in *AAAI Conference on Artificial Intelligence (AAAI)* (New York, NY: AAAI) 34, 12047–12054. doi: 10.1609/aaai.v34i07.6882

Szegedy, C., vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). "Rethinking the inception architecture for computer vision," in *Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV: IEEE), 2818–2826. doi: 10.1109/CVPR.2016.308

Tan, M., and Le, Q. V. (2019). "Efficientnet: rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning (ICML)*, Vol. 97 (Long Beach, CA: ACM), 6105–6114.

Too, E. C., Yujian, L., Njuki, S., and Yingchun, L. (2019). A comparative study of fine-tuning deep learning models for plant disease identification. *Comput. Electron. Agric.* 161, 272–279. doi: 10.1016/j.compag.2018.03.032

Toseef, M., and Khan, M. J. (2018). An intelligent mobile application for diagnosis of crop diseases in Pakistan using fuzzy inference system. *Comput. Electron. Agric.* 153, 1–11. doi: 10.1016/j.compag.2018.07.034

Touvron, H., Vedaldi, A., Douze, M., and Jegou, H. (2019). "Fixing the train-test resolution discrepancy," in *Neural Information Processing Systems (NeurIPS)*, MIT Press, Vancouver, Canada, 32, pp. 8252–8262.

Turkoglu, M., Hanbay, D., and Sengur, A. (2019). Multi-model LSTM-based convolutional neural networks for detection of apple diseases and pests. *J. Ambient Intell. Humaniz. Comput*. doi: 10.1007/s12652-019-01591-w. [Epub ahead of print].

Wu, S. G., Bao, F. S., Xu, E. Y., Wang, Y.-X., Chang, Y.-F., and Xiang, Q.-L. (2007). "A leaf recognition algorithm for plant classification using probabilistic neural network," *International Symposium on Signal Processing and Information Technology* (Giza: IEEE), 11–16. doi: 10.1109/ISSPIT.2007.4458016

Xiang, X., Yu, Z., Lv, N., Kong, X., and Saddik, A. E. (2020). "Semi-supervised image classification via attention mechanism and generative adversarial network," in *Proceedings Volume 11373, Eleventh International Conference on Graphics and Image Processing* (Hangzhou, China). doi: 10.1117/12.2557747

Xiao, F., Liu, B., and Li, R. (2020). Pedestrian object detection with fusion of visual attention mechanism and semantic computation. *Multimed. Tools Appl.* 79, 14593–14607. doi: 10.1007/s11042-018-7143-6

Yousefi, E., Baleghi, Y., and Sakhaei, S. M. (2017). Rotation invariant wavelet descriptors, a new set of features to enhance plant leaves classification. *Comput. Electron. Agric.* 140, 70–76. doi: 10.1016/j.compag.2017.05.031

Yu, Y., Jin, Q., and Chen, C. W. (2018). "FF-CMnet: a CNN-based model for fine-grained classification of car models based on feature fusion," in *International Conference on Multimedia and Expo (ICME)* (San Diego, CA: IEEE), 1–6. doi: 10.1109/ICME.2018.8486443

Zhang, L., Yang, Y., Wang, M., Hong, R., Nie, L., and Li, X. (2016). Detecting densely distributed graph patterns for fine-grained image categorization. *IEEE Trans. Image Process.* 25, 553–565. doi: 10.1109/TIP.2015.2502147

Zhang, S., Wu, X., You, Z., and Zhang, L. (2017). Leaf image based cucumber disease recognition using sparse representation classification. *Comput. Electron. Agric.* 134, 135–141. doi: 10.1016/j.compag.2017.01.014

Zhang, W., Tang, S., Su, J., Xiao, J., and Zhuang, Y. (2020). Tell and guess: cooperative learning for natural image caption generation with hierarchical refined attention. *Multimed. Tools Appl*. doi: 10.1007/s11042-020-08832-7. [Epub ahead of print].

Zhuang, P., Wang, Y., and Qiao, Y. (2020). Learning attentive pairwise interaction for fine-grained classification," in *AAAI Conference on Artificial Intelligence (AAAI)* (New York, NY: AAAI), 34, 13130–13137. doi: 10.1609/aaai.v34i07.7016

Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V. (2018). "Learning transferable architectures for scalable image recognition," in *Computer Vision and Pattern Recognition (CVPR)* (Salt Lake City, UT: IEEE), 8697–8710. doi: 10.1109/CVPR.2018.00907

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Yang, He, Yang and Xu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.