**Tomato Disease Classification and Identification Method Based on Multimodal Fusion Deep Learning**

Ning Zhang ¹,², Huarui Wu ¹,³,⁴,*, Huaji Zhu ¹,³,⁴, Ying Deng ¹,²,⁴ and Xiao Han ¹,³,⁴

¹ National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China
² College of Computer and Information Engineering, Beijing University of Agriculture, Beijing 102206, China
³ Research Center of Information Technology, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China
⁴ Key Laboratory for Quality Testing of Software and Hardware Products on Agricultural Information, Ministry of Agriculture and Rural Affairs, Beijing 100097, China
* Correspondence: wuhr@nercita.org.cn

**Abstract:** Considering that the occurrence and spread of diseases are closely related to the planting environment, a tomato disease diagnosis method based on Multi-ResNet34 multi-modal fusion learning based on residual learning is proposed for the problem of limited recognition rate of a single RGB image of a tomato disease. Based on the ResNet34 backbone network, this paper introduces transfer learning to speed up training, reduce data dependencies, and prevent overfitting due to a small amount of sample data; it also integrates multi-source data (tomato disease image data and environmental parameters). The feature-level multi-modal data fusion method is used to retain the key information of the data to identify the feature, so that the different modal data can complement, support and correct each other, and obtain a more accurate identification effect. Firstly, Mask R-CNN was used to extract partial images of leaves from complex background tomato disease images to reduce the influence of background regions on disease identification. Then, the formed image environment data set was input into the multi-modal fusion model to obtain the identification results of disease types. The proposed multi-modal fusion model Multi-ResNet34 has a classification accuracy of 98.9% for six tomato diseases: bacterial spot, late blight, leaf mold, yellow aspergillosis, gray mold, and early blight, which is higher than that of the single-modal model. With the increase by 1.1%, the effect is obvious. The method in this paper can provide an important basis for the analysis and diagnosis of tomato intelligent greenhouse diseases in the context of agricultural informatization.

**Keywords:** multimodal fusion; transfer learning; ResNet34; residual network; disease diagnosis

---

**1. Introduction**

Tomato is one of the important vegetable crops in China, and its planting area and total output rank first in the world. The increased occurrence of tomato diseases will seriously affect the yield and quality of tomatoes, such as tomato bacterial spot, late light, leaf mold, yellow leaf curl virus, Botrytis cinerea, early light, bacterial pulp necrosis, gray leaf spot, sclerotinia, etc. The traditional expert diagnosis efficiency is low and the number of experts is limited. Only relying on traditional expert knowledge and image processing methods to diagnose diseases cannot meet the development needs of agricultural informatization. However, crop disease surveillance is an important measure to ensure the healthy development of tomato industry. With the continuous development of computer vision [1], deep learning methods [2] have been widely studied and applied in the field of agricultural diseases. Many researchers have conducted a lot of studies in the field of image classification [3,4], and have achieved certain results. Some researchers use deep learning image recognition methods to carry out research. Lucas et al. [5] selected the AlexNet model to identify six kinds of apple disease images. The experimental results showed that the convolutional neural network method can achieve an accuracy of 97.3%. Guo Xiaoqing et al. [6]
proposed a multi-receptive field recognition model based on AlexNet for the problem of similar tomato disease identification. The average recognition accuracy of this model for early, middle and late tomato leaf diseases reached 92.7%. Wang et al. [7] compared de novo training and transfer learning to fine-tune the training model. The experimental results showed that transfer learning can effectively speed up the model convergence, and the accuracy rate was 90.4% on the VGG16 neural network. Wang Chunshan et al. [8] proposed a multi-scale residual lightweight Multi-scale ResNet disease identification model combined with a multi-scale feature extraction module in response to the large number of parameters and high computing and storage costs that occurred when the neural network model was deployed to agricultural Internet of Things equipment. The group convolution operation decomposed the large convolution kernel, the training parameters of the model reduced by about 93%, the overall size of the model reduced by about 35%, and the accuracy rate of 93.05% was obtained in the collected seven kinds of disease image data, which is a good result. At present, most of the research on vegetable disease identification at home and abroad is to identify a single disease image, since tomato diseases are closely related to the planting environment. Therefore, it is suggested to introduce multi-source information fusion to identify diseases and improve the quality of disease diagnosis.

With the vigorous development of wireless sensor networks and 5G transmission, agricultural greenhouses generate a large amount of agricultural production data, providing a data basis for intelligent disease diagnosis. Multi-source data reflects the growth status of crops from different perspectives. With the rapid growth of data, processing information and fusing different types of data are more and more important for crop disease diagnosis. With the development of deep learning and the rapid growth of multimedia data, and the fact that the actual problem scene is complex and changeable and contains various types of data, the research on multi-modal problems such as images, texts, and audios has been born [9,10]. Multimodal learning is composed of information from different modalities, usually including two or more modalities. The purpose is to combine the data of different modalities, pay attention to the internal relationship between modalities, and realize the mutual transformation of each modal information, even if it can fill in the missing information in the transfer process in the case of missing certain modalities. The data collected by the sensor has multi-modal characteristics, and the introduction of multi-modal information into disease identification helps to improve the accuracy of identification. Making full use of the complementarity and correlation between modalities to achieve multimodal data fusion has become a research focus in recent years [11–13]. Gao Ronghua et al. [14] collected environmental parameters in real time through the Internet of Things, integrated image feature fusion, environmental information and expert knowledge, and adopted the method of multi-structure parameter ensemble learning for disease identification to ensure the accuracy of identification under the condition of less identification time. In the experiment of 50 samples of four diseases of cucumber, the sample recognition rate was 79.4–93.6%. In order to enhance the accuracy and robustness of the visibility deep learning model under the condition of small samples, Shen Kecheng et al. [15] proposed a multi-modal visibility deep learning method based on visible light-far-infrared images, and constructed a multi-modal three-branch parallel structure. In the feature fusion network, the feature information of each branch realizes modal complementation and fusion through the network structure and outputs the visibility level corresponding to the image scene at the end of the network. Compared with traditional unimodal models, multimodal visibility models can significantly improve the accuracy and robustness of visibility detection under small sample conditions. Zhiqiang Yuan [16] learned by sharing the top hidden layer in each specific unimodal network Common representation of multimodal data. Wang et al. [17] demonstrated that a multimodal deep network structure can learn a good joint representation of audio and video data, thereby improving the recognition task accuracy. Chang Cheng [18] established a shared expression layer to fuse inter-modal information, and the improved model has stronger generalization ability and higher accuracy than the single-modal model.
As the crop growth environment will promote the occurrence of diseases, tomato diseases are closely related to the planting environment. The change of temperature and humidity in the planting environment will lead to the breeding of bacteria, which will lead to the occurrence and spread of diseases. The recognition accuracy is limited only by relying on a single tomato disease image data. Therefore, multi-source information fusion is introduced to identify diseases and improve the quality of disease diagnosis. The method of disease identification based on multimodal fusion uses the complementarity of data to improve the accuracy of disease identification. Based on the research results of convolution neural network theory and the characteristics of tomato leaf surface disease images, this paper will study the multi-mode fusion tomato disease recognition model based on migration learning [19–21] for disease images in small data domain and environmental data and explore multi-mode data fusion analysis and processing methods. The accuracy of tomato disease identification is improved through multi-modal information recognition method. This paper considers the environmental factors in the process of disease occurrence, and analyzes the disease types in combination with disease images and environmental factors when the disease occurs, which makes the tomato disease recognition model closer to the actual production scene and provides research ideas for vegetable disease diagnosis.

2. Dataset and Methods

2.1. Dataset

2.1.1. Data Sources

The dataset in this paper is composed of tomato leaf disease RGB (R stands for Red, G stands for Green, and B stands for Blue) image dataset and environmental dataset, and a one-to-one correspondence between image data and environmental data is established. Among them, the tomato disease image dataset includes 2 types of data. Data 1 comes from the leaf surface images of tomato diseases in Plant Village, including bacterial spot, late blight, leaf mold, yellow smut and healthy leaves, each with 800 images. There are a total of 4000 images, and the samples are simple background leaf surface disease images taken in the laboratory environment. Data 2 is obtained from the tomato greenhouse facility experimental site of the tomato greenhouse facility in the Changping District, Beijing, in the Xiaotangshan National Precision Agriculture Research Demonstration Base. Diseased leaf surface images, including leaf mold, botrytis, late blight, and early blight, were collected under natural light conditions, each with 300 images, totaling 1200 images. The samples are images of leaf surface diseases with complex backgrounds taken in actual production environments. These two parts of tomato disease images together constitute the experimental image dataset in this chapter. Images of six tomato diseases are shown in Figure 1.

The environmental parameter data set includes the data collected by sensors in the tomato greenhouse facility of the experimental base and the simulated data generated according to the disease occurrence conditions reviewed in the literature. Since the occurrence of disease is related to environmental temperature and humidity parameters, air temperature and air humidity are selected as environmental parameters. The relationship between disease and ambient temperature and humidity is shown in Table 1. The sensors collected environmental data of diseases and screened 3200 pieces of environmental data of each kind of disease.

2.1.2. Extract Diseased Leaves

In order to avoid the influence of non-leaf disease areas on tomato disease identification which leads to a decrease in the accuracy of disease identification, the dataset images with complex backgrounds are first processed. The image segmentation technology was used to extract the diseased leaf surface area in the image, and the Mask R-CNN convolutional neural network was selected to process the data. After uniformly scaling the complex background disease images, Label Me was used to label the leaf surface area and background area of tomato disease. The annotated image data is input into the Mask
R-CNN network for processing, and the non-leaf area of the tomato disease image is filled with zero pixels, that is, the pixel value of the tomato diseased leaves is retained, and the rest filled with zero pixels. The complex background image processing flow is shown in Figure 2 below.

Table 1. The relationship between disease, temperature and humidity.

<table>
<thead>
<tr>
<th>Type of Disease</th>
<th>Onset Temperature/°C</th>
<th>Onset Humidity/%rh</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tomato bacterial spot</td>
<td>10–20</td>
<td>85–100</td>
</tr>
<tr>
<td>Tomato late blight</td>
<td>20–30</td>
<td>90–100</td>
</tr>
<tr>
<td>Tomato leaf mold</td>
<td>20–25</td>
<td>80–100</td>
</tr>
<tr>
<td>Tomato yellow leaf curl virus</td>
<td>25–33</td>
<td>25–45</td>
</tr>
<tr>
<td>Tomato Botrytis cinerea</td>
<td>20–25</td>
<td>85–100</td>
</tr>
<tr>
<td>Tomato early blight</td>
<td>25–32</td>
<td>85–100</td>
</tr>
</tbody>
</table>

Figure 1. Example of Tomato Diseased Leaf Data Set. (a) Tomato leaf image—simple background image taken in the laboratory, 256 × 256 pixels. The image types include Tomato Leaf mold, Tomato Late night, Tomato normal leaves, Tomato Vertical spot, Tomato Yellow leaf curl virus. (b) Tomato leaf image—complex background image taken in greenhouse, 3024 × 3024 pixel. Image types include Tomato Leaf mold, Tomato Late blind, Tomato normal leaves, Tomato Botrytis cinerea, and Tomato Early blind.
2.1.3. Data Augmentation

In order to prevent overfitting of the network, enrich the number of images, and simulate the actual shooting of disease images under different lighting conditions, shooting angles, and leaf deformation to enrich the image types, the images are enhanced by increasing the image brightness, reducing the image brightness, flipping the image horizontally, and rotating the image. Six data enhancement methods of 15 degrees, image rotation 30 degrees, and image stretching are used for processing. The data enhancement display is shown in Figure 3. After the expansion, each type of disease-environment data reaches 3200.

2.2. Model
2.2.1. ResNet34 Backbone Network

Since the convolutional neural network AlexNet shines in the 2012 ImageNet competition, many excellent convolutional neural networks have been proposed by researchers for image classification tasks. The researchers improved the performance of the model by increasing the network depth, increasing the function of the convolution module, adding new functional units and facing the tasks in the field of complex image recognition, and derived many convolutional neural networks with excellent results, such as VGG, GoogleNet, Fast R-CNN, SeNet, etc. A deeper and wider network architecture means that the feature information that the model can extract is richer and more semantic. However, as the number of network layers deepens, problems such as gradient disappearance, gradient explosion, and network degradation occur. Among them, the ResNet neural network model uses the residual module to effectively solve the problem of network degradation after the network is deepened. Tomato disease spots have the characteristics of large intra-class differences and small inter-class differences, and the color and size of disease spots of different species and different periods are different. Some disease spots are scattered, and it is difficult to determine the disease type. In this paper, the accuracy of tomato disease classification is improved by increasing the depth of the model. In order to avoid the degradation of
the gradient vanishing model caused by the increase of depth, ResNet34 is used as the backbone model. The residual network avoids model degradation and causes the network to converge by superimposing the residual modules directly connected across layers for many times while increasing the depth of the network. The residual connection in the network is a special mapping, and the residual structure is shown in Figure 4 below. The residual module is very important to the residual network, and the residual model can be formed by continuously stacking the residual modules. The left side of the residual module is an ordinary convolutional network structure, and the right side is directly connected. It helps the data to learn the identity mapping by means of skip connections, so that the output of the shallow layer and the output of the deep layer are summed as the input of the next layer. The input to the ResNet34 model is a 224 × 224 three-channel image with a total of 34 layers. The initial layer of the model is an ordinary convolution structure. After the first layer of 7 × 7 convolution images, the dimension is reduced to 112 × 122, and the number of channels is increased to 64. After the maximum pooling, the image dimension is reduced to 56 × 56, and the number of channels remains unchanged. Next, the residual part occurs. The structure with multiple residual units is called layer, which contains 4 layers in total. The size of the convolution kernel of the residual module is 3 × 3. After each layer, the residual image is reduced in dimension. Half the number of channels is increased to 2 times of the original. After 4 layer operations, the image dimension is reduced to 7 × 7, and the number of channels is increased to 512; finally, the average pooling layer and the fully connected layer are connected.

![Figure 4. Residual module structure.](image)

2.2.2. Transfer Learning

The occurrence of tomato diseases in agricultural operation environment is difficult to predict, and it is difficult to collect and label tomato image data, and the actual number of effective tomato disease images is small. Aiming at the problems of model overfitting and data outliers that are easily caused by image recognition in small data domains, on the one hand, the data enhancement method is used to increase the number of images; on the other hand, the transfer learning method is used to accelerate the training convergence speed and avoid model overfitting. Transfer learning is a machine learning method that uses existing knowledge to solve problems in related fields. Deep transfer learning has many advantages, such as reducing data dependencies, speeding up training, and improving training efficiency. In this paper, the parameters of the ResNet34 model pre-trained on ImageNet are transferred to the tomato leaf disease image data domain and in the built
Muti-ResNet34 multimodal fusion model, the parameters of the image feature extraction layer are fixed, and the fully connected layer and softmax of the model are changed. The fully connected layer is rebuilt for 7 classifications and then trained on the improved part. The training process of ResNet model parameter transfer learning is shown in Figure 5. The fully connected layer Adam acts as the optimizer and selects the cross entropy function to train the loss value. The formula is as follows:

\[
L = H(p, q) = - \sum_i (p(x) \log q(x)).
\]  

(1)

Figure 5. The training process of tomato disease image classification model based on transfer learning.

In the formula, \(H(p, q)\) represents the cross entropy; the probability distribution \(p\) is the expected output, and the probability distribution \(q\) is the actual output.

2.2.3. Multimodal Data Fusion Methods

For tomato disease identification, the RGB disease image modal contains rich disease color and texture information, and the environmental temperature and humidity parameter modal reflects the disease occurrence from the perspective of the spatial temperature and humidity information where the disease is located. This paper considers the complementarity between the information. In this paper, a tomato disease identification method based on the above two modal characteristics was proposed. According to the division method of data fusion level, data fusion is usually divided into data-level fusion, feature-level fusion and decision-level fusion. The data level fusion fuses the information of different modalities before identification and retains as much useful information as possible, but due to the large amount of information, it takes a long time and the data stability is poor; the feature level fusion features the information of different modalities. After extraction, fusion is performed, and then the feature vectors are classified into meaningful combinations, and redundant information is compressed. The fusion result can maximize the feature information required for decision analysis. After independently completing the decision-making task, the decision-making results are fused, and the joint decision-making result is output, which can process asynchronous information and effectively fuse the information of different dimensions and types of the same task. Since the feature level fusion can provide the maximum feature information required for decision analysis, the feature level fusion is performed on the backbone model. Based on the image diagnosis of disease types, environmental information is introduced, and a more accurate decision basis for disease identification is obtained through the mutual complementation of multi-modal data. The diseased leaf image data and environmental data are fused at the feature level, and the fusion results are classified. The original data is transformed into a high-dimensional feature representation by using a neural network, and then the feature data extracted from different modalities are fused in the high-dimensional space. Firstly, the weights obtained by ResNet34 pre-training on the ImageNet dataset are used for feature extraction of tomato disease leaf images through transfer learning. To avoid adverse effects, the convergence speed is improved, and the environment data normalized to obtain the environment vector; then, the Concat fusion operation is performed on the extracted image features and the processed environment vector, Concat fuses feature vectors by increasing the number of
channels, and the fusion result is inputted directly into the fully connected layer. The operation process of the Concat layer is as follows:

\[ S_c = f_{\text{concat}}(M_i, K_i). \] (2)

In the formula, \( S_c \) is the feature after fusion; \( f_{\text{concat}} \) is the feature fusion operation; \( M_i \) is the disease image feature vector; \( K_i \) is the environment feature vector.

2.2.4. Residual Network-Based Multimodal Fusion Method

The architecture based on the deep convolutional neural network has achieved excellent performance in image modeling. In this paper, the ResNet34 convolutional neural network is selected as the backbone model, and the tomato bacterial spot, late blight, leaf mold, yellow aspergillosis are examined. Taking typical diseases such as gray mold and early blight as examples, the deep learning convolutional neural network method of multi-modal data fusion is studied, and the Multi-ResNet34 multi-modal fusion model is proposed. The disease identification algorithm framework based on multi-modal feature fusion is shown in the Figure 6. The model extracts image features and temperature and humidity parameter features from the input layer “image of tomato diseased leaves, air temperature and air relative humidity parameters”, two different structure data decibels through two paths, and fuses feature weights to obtain fusion features of multiple structural parameters. The device judges the disease type and obtains the identification result. The disease image data and environmental parameters are combined to realize tomato disease identification. The feature-level multimodal data fusion method is used to retain the key information of the identifiable features of the data. The scheme is identified.

![Tomato Multimodal Data Fusion Recognition Multi-ResNet34 Model](image6)

**Figure 6.** Tomato Multimodal Data Fusion Recognition Multi-ResNet34 Model.

3. Experimental Results and Analysis

3.1. Test Environment and Parameter Settings

Test environment: Windows10 system, processor Intel(R) Core(TM) i5-5200 U CPU@2.20 GHz 2.19 GHz, RAM 8.0 GB, Python 3.6, Pytorch 1.2.0, JetBrains PyCharm 2018.3.3 ×64.

The multimodal image environment dataset is composed of seven types of image data and environment data. There are 3200 pieces of disease environment data for each type, totaling 22,400 groups of data. The dataset is divided into training set and test set at a ratio of 7:3. The Torch deep learning open-source framework is used to build the network model, ImageNet is used to pre-train weights, the optimizer selects the computationally
efficient Adam optimizer, and the loss function selects the commonly used cross-entropy loss function. The learning rate is set to 0.0001; batch training is used to divide the training set and test set into multiple batches, and the number of samples selected for each batch of training is 16, that is, the train batch is set to 16; epoch means all the training set samples. After completing a forward propagation and back propagation, the epoch determines the number of times the network needs to be trained on all samples of the training set, and the epoch is set to 30.

3.2. Evaluation Indicators

There are various evaluation indicators for evaluating the performance of the model. Through different evaluation indicators, the recognition effect of the model can be fully understood from different angles. Benben adopts the following two evaluation methods to evaluate the performance of the model.

(1) The average recognition accuracy (AA) can measure the recognition accuracy of the model to the disease, and it is convenient to compare the recognition performance of different models from a macro perspective. The average recognition accuracy $P$ refers to the ratio of correctly predicted samples to all observations in the model, as shown in the following formula:

$$P = \sum_{i=1}^{M} \frac{TP_i}{TP_i + FP_i} \times 100\%.$$  

In the formula, $TP_i$ represents the number of correctly predicted samples of the i-th disease sample in the test set, $FP_i$ represents the number of samples that are incorrectly predicted by the i-th disease sample in the test set, and $M$ represents the number of disease types.

(2) The confusion matrix clearly shows the number of correctly classified and misclassified classes, as well as the number of each feature. The overall classification accuracy and Kappa coefficient were calculated from the confusion matrix to determine the accuracy and reliability of the evaluation results. The overall classification accuracy (OA) can well represent the accuracy of the model classification. Kappa coefficient is used to measure the classification accuracy of the model and is an indicator for consistency test. Consistency refers to whether the predicted results of the model are consistent with the actual classification results. The calculation result is usually between 0 and 1, and the closer the value is to 1, the higher the consistency. The Kappa coefficient and overall classification accuracy are calculated as follows:

$$P_o = \frac{C}{N},$$  

$$P_e = \frac{\sum_{i=1}^{M} a_i b_i}{N^2},$$  

$$k = \frac{P_o - P_e}{1 - P_e}.$$  

Among them, $P_o$ is the overall classification accuracy; $C$ is the number of correctly classified samples for each category; $M$ is the number of disease types; $N$ is the total number of samples; $a_i$ is the number of real samples in each category; $b_i$ is the predicted sample for each category number; $k$ represents the Kappa coefficient.

3.3. Multi-ResNet34 Model Recognition Results

The training time of Multi ResNet34 model in tomato disease image environment dataset is 3068 s. Through the model testing, the classification of the environmental data of seven kinds of tomato disease images, the obtained seven-category confusion matrix
(Confusion Matrix) and the recognition rate of each type of disease are shown in Figure 7 and Table 2 below. The confusion matrix clearly shows the misclassification of each type of disease. As can be seen from the figure, most of the data is classified correctly. The overall classification accuracy of the model is 97.3%, and the Kappa coefficient is 0.968. The overall judgment is accurate and the recognition rate is high. Among them, the identification accuracy of tomato late blight was the lowest at 95.00%. Because the late blight and gray mold lesions both showed large gray-brown lesions and occurred in a low temperature and high humidity environment, the two were most likely to be misjudged as each other.

![Confusion Matrix](image)

**Figure 7.** Confusion matrix of recognition results. 0. Tomato bacterial spot; 1. Tomato healthy; 2. Tomato late blight; 3. Tomato leaf mold; 4. Tomato yellow leaf curl virus; 5. Tomato Botrytis cinerea; 6. Tomato early blight.

**Table 2.** The recognition rate of each type of disease image on the multi-modal.

<table>
<thead>
<tr>
<th>Type of Disease</th>
<th>Number of Test Samples</th>
<th>Accuracy/%</th>
<th>Recall/%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tomato bacterial spot</td>
<td>100</td>
<td>98.00</td>
<td>98.00</td>
</tr>
<tr>
<td>Tomato healthy</td>
<td>100</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Tomato late blight</td>
<td>100</td>
<td>95.00</td>
<td>96.94</td>
</tr>
<tr>
<td>Tomato leaf mold</td>
<td>100</td>
<td>97.00</td>
<td>95.10</td>
</tr>
<tr>
<td>Tomato yellow leaf curl virus</td>
<td>100</td>
<td>97.00</td>
<td>98.98</td>
</tr>
<tr>
<td>Tomato Botrytis cinerea</td>
<td>100</td>
<td>96.00</td>
<td>96.97</td>
</tr>
<tr>
<td>Tomato early blight</td>
<td>100</td>
<td>98.00</td>
<td>95.15</td>
</tr>
</tbody>
</table>

### 3.4. Comparative Experiment

In order to verify the effectiveness of the method in this paper, three aspects are considered: the selection of the backbone model, the location of multi-structure data fusion and the addition of environmental data. The comparison experiments are MobileNetV2, InceptionV3, ResNet34, ResNet54 decision fusion, Multi-ResNet34. Including six groups of experiments, each model is iterated 30 times, and the generalization and fitting ability of each model is judged by whether the loss function curve converged, combined with the performance of each model on the test set; the accuracy is compared. Table 3 shows the experimental results of different models. Overall, the disease recognition model based on ResNet34 multimodal feature fusion proposed in this paper is better than the baseline model.

Comparing the training process of a single image dataset of different backbone networks, the accuracy rates of the three groups of models of ResNet34, MobileNetV2, and InceptionV3 on the test set are 97.8%, 94.3%, and 95.8%, respectively. The curve changes of training accuracy (accuracy) and cross entropy (cross entropy) during the training of
ResNet, MobileNetV2 and Inception V3 networks are shown in part A in Figure 8, respectively. It can be seen from the curve that during the training process of ResNet34, the accuracy increases and the cross entropy decreases rapidly, and the curve does not fluctuate much. After 10 iterations or so, the curve tends to be stable and achieves a high accuracy rate; MobileNet convergence iteration speed is slower than that of ResNet34. Iteration tends to be stable after about 20 iterations, and the final accuracy rate is the lowest among the three groups; because the InceptionV3 network is wider and deeper, the calculation amount is large, the increase in the accuracy of InceptionV3 and the decrease in cross entropy are relatively slow, and the time required to achieve stability is longer than that of other networks. The training curve fluctuates the most. Therefore, this paper proposes ResNet34 as the backbone network.

Table 3. Comparison of the recognition accuracy of different models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset</th>
<th>Accuracy/%</th>
</tr>
</thead>
<tbody>
<tr>
<td>MobileNetV2</td>
<td>Tomato disease image dataset</td>
<td>94.3%</td>
</tr>
<tr>
<td>InceptionV3</td>
<td>Tomato disease image dataset</td>
<td>95.8%</td>
</tr>
<tr>
<td>ResNet34</td>
<td>Tomato disease image dataset</td>
<td>97.8%</td>
</tr>
<tr>
<td>ResNet34 Decision Fusion</td>
<td>Environmental image data set of tomato diseases</td>
<td>98.1%</td>
</tr>
<tr>
<td>Multi-ResNet34</td>
<td>Environmental image data set of tomato diseases</td>
<td>98.9%</td>
</tr>
</tbody>
</table>

Figure 8. The training process of different models on the tomato disease dataset.

Compared with the single tomato disease ResNet34, the accuracy of disease diagnosis was improved after adding environmental data. It can be seen from the curve that both the single-modal and multi-modal models converge faster, and the curve of the Multi-ResNet34 model has less fluctuation and higher stability. By comparing and verifying the superiority of feature fusion according to different fusion positions, a total of two schemes of ResNet34 decision fusion and Multi-ResNet34 are set up for comparison experiments.
Among them, the KNN model is used in the ResNet34 decision fusion model to train the environmental data, and the output results of the environmental model and the image model are fused at the decision level, and the predicted value of the text model and the predicted value of the single image model are averaged. The accuracy rates of the two models of ResNet34 decision fusion and Multi-ResNet34 on the test set are 98.1% and 98.9%, respectively. Observing the model training process in part B in Figure 8, it can be seen that the Multi-ResNet34 model is more stable during training. This is because the complementarity between the model data improves the accuracy of the overall disease diagnosis. The image environment feature fusion model proposed in this paper considers the key information that affects the occurrence of diseases, the temperature and humidity information of the planting environment, uses ResNet34 to mine the disease information in the disease image, and fuses the environmental features and image features before making a decision. The recognition accuracy on the test set is 97.8%; the multimodal model has a recognition accuracy of 98.9% on the test set, and the accuracy is increased by 1.1%. The experimental results show that the introduction of the environmental temperature and humidity data feature fusion algorithm can significantly improve the accuracy of crop disease identification.

4. Conclusions

Timely identification of diseases is very important for agricultural production personnel. It can help agricultural practitioners to identify disease types in a timely manner, reduce the time required to consult experts on disease types, identify disease types more quickly and respond, and reduce economic losses caused by crop diseases. It is of great significance for agricultural production. The closed greenhouse and repeated planting of crops lead to overlapping of greenhouse crop diseases, which increases the difficulty of disease identification. Aiming at the problem that the accuracy of single modal data is limited in disease diagnosis, this paper proposes a disease recognition algorithm based on deep learning multi-modal feature fusion Multi-ResNet34 convolutional neural network model. Differences and complementarities improve the accuracy of disease diagnosis. The self-built bacterial spot disease, late blight, leaf mold, yellow smut, gray mold, and early blight image environment dataset of six types of tomato diseases are preprocessed by image segmentation and data enhancement methods to improve image recognition, efficiency, and increase the amount of data. The improved Multi-ResNet34 model has an accuracy of 98.9% in the classification of six tomato diseases, which is 1.1 percentage points higher than that of the single image recognition model, providing a certain reference for the diagnosis of greenhouse crop diseases under agricultural big data.

Author Contributions: N.Z.: Conceptualization, Data curation, Software, Validation, Visualization, Writing—original draft. H.W.: Investigation, Formal analyze. H.Z.: Data curation. Y.D.: Data management. X.H.: Writing, Reviewing and Editing, Supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by the National Key R&D Program (Approval No.: 2020YFD1100602), supported by the National Modern Agricultural Industry Technology System (CARS-23-C06), and funded by the Youth Science Fund of Beijing Academy of Agriculture and Forestry Sciences (Approval No.: QNJJ202030). The authors thank the funding agencies for their financial support.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Thank the above funds for their financial assistance to this article.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
References


