Computers and Electronics in Agriculture 189 (2021) 106408

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Crop disease identification and interpretation method based on multimodal deep learning

Ji Zhou [a,b,e], Jiuxi Li [d,e,*], Chunshan Wang [a,b,c,e,*], Huarui Wu [b,c], Chunjiang Zhao [b,c], Guifa Teng [a,e]

[a] School of Information Science and Technology, Hebei Agricultural University, Baoding 071001, China
[b] National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China
[c] Beijing Research Center for Information Technology in Agriculture, Beijing 100097, China
[d] School of Mechanical and Electrical Engineering, Hebei Agricultural University, Baoding 071001, China
[e] Hebei Key Laboratory of Agricultural Big Data, Baoding 071001, China

ARTICLE INFO

Keywords:
Multimodality
Knowledge graph
Image-text pairs
Semantic interpretation
Disease identification

ABSTRACT

Identification methods of crop diseases based on image modality alone have achieved relative success under limited and restricted conditions. As a data-driven technology, its performance depends on a large amount of image labeling data. Many of the existing methods neglected the role and value of other modal data except images and only relies on low-level image features for disease recognition without utilizing high-level domain knowledge, leading to poor credibility and interpretability of identification results. This paper targets tomato and cucumber common invasive diseases recognition as the research object. First, for the problem of insufficient utilization of multimodal data in existing models, the semantic embedding methods for disease images and disease description texts were examined, and the correlation and complementarity between the two types of modal data was utilized to realize the joint representation learning of disease features. Second, in response to the requirements of reliable identification and interpretability of diseases, the knowledge representation and knowledge embedding mechanism in the field of disease identification was studied, and the high-level domain knowledge graph was used as the external guidance for image feature learning and disease identification. Lastly, a disease identification model based on "image-text" multimodal collaborative representation and knowledge assistance (ITK-Net) was constructed. The proposed model achieved an identification accuracy, precision, sensitivity and specificity of 99.63%, 99%, 99.07% and 99.78% respectively on a dataset composed of "image-text" pairs. Meanwhile, semantic interpretation was performed on the model inference process. The achievement of this paper can offer a new method for disease identification based on multimodal data and domain knowledge, which might help improve the intelligence level of crop disease identification.

1. Introduction

Crop diseases are one of the most critical challenges that have long impacted the safety of agricultural production, agricultural products and ecological environment. Such diseases not only severely affect the yield and quality of agricultural products but also the overall benefits of the agricultural industry. According to the estimates of the Food and Agriculture Organization of the United Nations, the average loss of crops caused by diseases was around 10–30% of the total yield (Savary et al., 2019). Therefore, rapid and accurate identification of crop diseases is the first step to adopt effective prevention and control measures to stop such losses in time. With the development of information science, cutting-edge technologies such as image processing and machine learning have been widely used in the diagnosis of crop diseases, and have provided powerful techniques and approaches for rapid, accurate and non-destructive disease identification.

Ferentinos (2018) carried out model training using an open database of 87,848 images covering 25 different plants in 58 [plant, disease] combinations of different categories, and achieved an identification accuracy of 99.53%. Ma Juncheng et al. (2013), proposed a Deep Convolutional Neural Network (DCNN) for identifying cucumber anthracnose, downy mildew, powdery mildew and target leaf spot disease. On

* Corresponding authors at: Hebei Agricultural University, Baoding 071001, China (C. Wang), (J. Li).
E-mail address: chunshan9701@163.com (C. Wang).

https://doi.org/10.1016/j.compag.2021.106408
Received 26 March 2021; Received in revised form 23 June 2021; Accepted 14 August 2021
Available online 25 August 2021

an extended dataset containing 14,208 symptom images, this DCNN achieved an identification accuracy of 93.4%. Picon et al. (2019), proposed an adapted Deep Residual Neural Network-based algorithm to realize early identification against Septoria, Tan Spot and Rust, and achieved an identification accuracy of 96%. Although the aforementioned studies have obtained satisfactory identification results on specific datasets, the main shortcomings of relying solely on image modal data for disease representation learning should not be neglected. On the one hand, the image modal data cannot cover all the features of a disease and needs to be supplemented by data of other modalities; on the other hand, the model can only learn low-level image features, and the features that the identification decision relies on are difficult to understand. Since the results of disease identification directly affect the subsequent prevention strategy and drug spraying, interpretability as a sensitive task has become a difficult problem in the development and application of deep learning in the field of disease identification. With regards to the various concerns and challenges related to the application and interpretability of multimodal data, a number of scholars have conducted fruitful research from different perspectives.

In terms of application of multimodal data, Frome et al. (2013), solved the classification problem of multi-label values in a better way by fusing image modal data with text modal data, and claimed that this method was effective for zero-sample learning. In order to improve the identification precision of zero-sample fine-grained images, Akata et al. (2015), combined image and text modal information and achieved remarkable results. Xu et al. (2018), embedded the structured knowledge database and unstructured images and texts into the network model, which improved the identification precision of fine-grained images and the interpretability of classification results. In addition, Peng et al. (2019) and He and Peng (2017) embedded the image modal and text modal data separately and jointly to further improve the classification accuracy of fine-grained images. Inspired by the studies above, Zhong et al. (2020), solved the problem of absence or lack of certain disease features in disease images in the learning process of zero-sample and small-sample disease identification by adopting both the image modal and text modal data, supplemented with noise information. Therefore, the application of multimodal data for fine-grained image classification and zero-sample learning has achieved good results and shown a high level of interpretability.

In terms of model interpretability, most of the existing studies focused on visualization. Ghosal et al. (2018), constructed an interpretable neural network model and realized the identification, classification and quantification of plant stress. Erhan et al. (2009) and Simonyan et al. (2013) performed gradient ascent in the feature map activation part and image data input part respectively, and then further explored the internal working mechanism of the neural network. Tobias et al. (2014), proposed a neural network that was fully constituted by convolutional layers and introduced a guided back propagation mode. Smilkov et al. (2017), proposed a method to determine the key area in the image by adding noise into the input image and observing the gradient change of the image after adding the noise. Chattopadhay (2018) and Selvaraju et al. (2017) explained the mechanism of convolutional neural network in the form of a heat map, and then indirectly explained the focus area of the model. Xu et al. (2020):, further improved the interpretability of the visualization results by eliminating the baseline and removing some visual artifacts that are easy to appear in the interpretation. Dai (2020), discussed model interpretability from two perspectives, i.e., visualization of the convolution process and hierarchical clustering. Lai et al. (2020), examined the mapping relationship between human attention and the attention model designed by neural network, and carried out experiments and explanations in three aspects, i.e., target segmentation, video action identification, and fine-grained image identification.

In terms of the process of disease identification, an increasing number of studies are tending to simulate the human process in disease observation and identification. This approach can not only improve the performance of the model, but also provide a kind of semantic interpretation for the model. Therefore, in order to improve the identification accuracy and interpretability of the disease identification model, the present study built a knowledge graph on specific tomato and cucumber diseases, and established a disease identification model based on "image-text" multimodal collaborative representation and knowledge assistance (ITK-Net) by utilizing the image modal features, text modal features, knowledge graph modal features and their respective feature learning methods. The main contributions of this research are as follows:

1)A knowledge graph for specific crop diseases was established to provide knowledge assistance for the disease identification process.

2)A crop disease identification model based on "image-text" multimodal collaboration and knowledge assistance was established by fusing unstructured data (image modality and text modality) and structured data (knowledge graph modality).

3)In the process of disease identification, the inference process was simulated in different ways so as to interpret the model from a semantic perspective.

The remainder of the paper is organized as follows: In Section 2, the detailed information is provided, including the data set, the construction method of the knowledge graph, the ITK-Net model and metrics used in our experiment. Section 3 contains the comparison of the control group model and the visualization of different modal models. We discusses the limitations of this research method in Section 4. Finally, we provide a conclusion on the proposed approach in Section 5.

2. Materials and methods

2.1. Data acquisition

All the datasets used in this study were collected from the Xiaotangshan National Precision Agriculture Demonstration Base. The self-collected disease image dataset was composed of a total of 1,715 images and 1,715 text records covering 6 types of diseased leaves (tomato powdery mildew, tomato early blight, cucumber powdery mildew, cucumber virus disease, cucumber downy mildew, and cucumber bacterial spot), which were divided into the training set, the validation set and the test set in accordance with the ratio of 7:2:1. The images of the diseased leaves were captured in three time periods: morning (7:00–8:00), noon (11:00–12:00) and evening (17:00–18:00). The description text data was manually provided by 5 plant protection experts. Table 1 shows examples of the dataset.

2.2. Introduction to common diseases of tomato and cucumber

In this study, six typical diseases of tomato and cucumber crops were selected as the research objects. The common feature of these diseases is that they are prone to occur and spread in most greenhouse planting environments in China, especially in North China, and have greater harm to yield and quality. Studies have shown that temperature and humidity are important conditions leading to the occurrence of the above-mentioned diseases. The suitable temperature for the cucumber bacterial spot is 18℃-26℃, and the suitable relative humidity is 75%. The temperature of cucumber downy mildew epidemic is 20℃-24℃, and the relative humidity is greater than 85%. Especially when there is a water film on the leaves, it is more conducive to swim with the spores and invade the host. When the temperature is 15℃ and the relative humidity is greater than 80%, tomato early blight is prone to occur. The optimum temperature for tomato powdery mildew is 22℃-28℃, and the relative humidity is 40–95%. Cucumber virus disease is mainly transmitted by insects. Under high temperature and arid environment, it is conducive to insect reproduction and disease transmission.

2.3. Construction of disease knowledge graph

The crop disease knowledge graph can provide knowledge assistance

Table 1
Examples of the dataset.

Disease category	Number of "image-text" pairs	Image	Description text
Tomato powdery mildew	238		Several white oval spots are distributed on the front side of tomato leaves.
Tomato early blight	285		The front side of tomato leaves is fully covered with oval yellow–brown ring spots.
Cucumber powdery mildew	299		The front side of cucumber leaves is fully covered with white dot-like spots.
Cucumber virus disease	363		There are yellow-green patterns on cucumber leaves, and the edges of the leaves are wrinkled.
Cucumber downy mildew	331		There are several polygonal yellow spots on the front side of cucumber leaves.
Cucumber bacterial spot	299		Yellow-white irregular spots are densely distributed on the front side of cucumber leaves.

for the disease diagnosis process. Therefore, it is critically important for establishing a knowledge graph with comprehensive coverage and accurate description. In this study, by crawling knowledge descriptions from Baidu Encyclopedia and other websites about tomato and cucumber disease prevention and control, a disease knowledge graph was established based on triples as the basic unit following the procedure of entity identification, attribute relationship establishment and attribute value extraction. The specific construction process of the disease knowledge graph is shown in Fig. 1. In the entity recognition part, the original disease description text is processed by word segmentation (the word segmentation tool uses jieba), and the word after segmentation is obtained to determine its part of speech, and non-substantive words such as conjunctions are removed, and substantive words are retained as entities in the knowledge graph. The attribute relationship establishment part adopts manual labeling method. Under the guidance of plant protection experts, attribute relationships are established according to the occurrence characteristics of different crops. The attribute value extraction part is based on the pre-defined disease attribute relationship. Among all entities in the original text, disease category entities (such as tomato powdery mildew, cucumber downy mildew and other entities) are excluded, and the remaining entities and attribute relationships constitute a complete disease knowledge triplet.

2.4. Vectorization of description text and knowledge graph

In this study, the disease description text and knowledge graph were both constructed in the Chinese language. In the training process of the multimodal identification model, the description text and knowledge graph were vectorized first. After considering the parameter consumption in training, the bag of words model was used to represent the disease description text, and the maximum value of representation vector for each description sentence was limited to 20 (omitted if exceeding the maximum value; complemented with 0 if insufficient). The word2vec model was used to represent the knowledge graph. It is possible to make the distance between similar attribute values closer and make the distance between attribute values with large semantic differences farther by using sufficient vectors for text representation. Thus, the representation vector dimension of each attribute was set to 100. The text vectorization process is shown in Fig. 2.

2.5. Construction of the ITK-Net

The image modality provides visual information for the disease identification model in the form of dense vectors, the text modality provides text information for the model in the form of sparse vectors, and the knowledge graph provides knowledge information for the model in the form of structured graph data. The image modality and text modality adopt the convolutional neural network and cyclic neural network for input feature learning, respectively. After obtaining the identification model combining the "image-text" modalities, the knowledge was then added to assist the diagnosis process during disease inference and analysis. The specific model structure is shown in Fig. 3.

The image modal data and text modal data are unstructured data. During the training process, the feature extraction of both data was mainly achieved by the convolution kernel window. The knowledge graph data is graph data (that is, structured data), and its node search method adopts jump search between nodes. Therefore in this study, the model training and model application were implemented by applying two different strategies.

2.5.1. Model training process

The input data of the image modality contains rich visual features (e.g., the color, shape and location of disease spots), which are represented by T_i^{img}; the input data of the text modality contains image visual features in the form of text expression, which is represented by T_i^{text}. The visual features of the disease in the images are extracted by the deep convolutional neural network. Compared with image features, the dimension of text features is greatly reduced, but the semantic representation of a single text feature is abundant. Therefore, the recurrent neural network was used to extract the features of a single input vector and its surrounding features. The joint expression of image features and text features is shown in Eq (1).

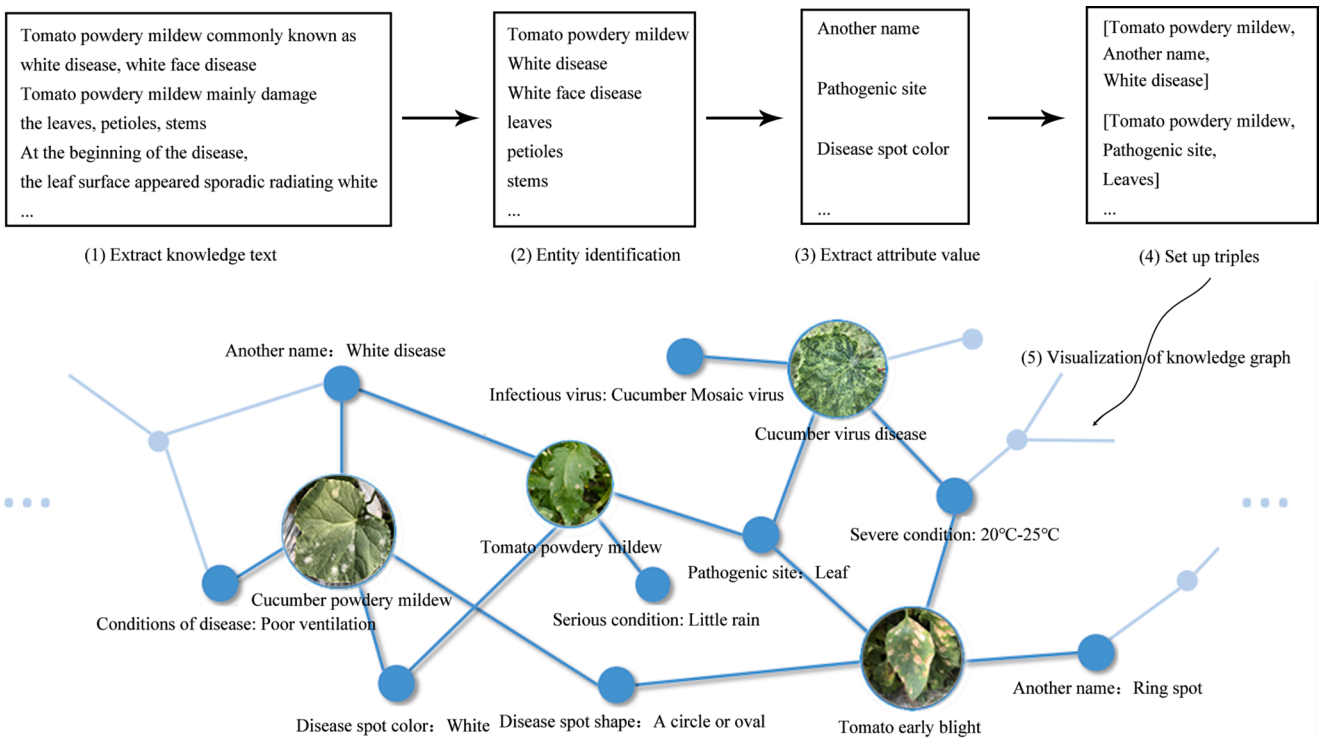

Fig. 1. The construction process of the disease knowledge graph.

$$P_{i\&t} = \sum_{i=0}^{N} C\left(T_i^{img}, L\right) + R\left(T_i^{text}, L\right) \tag{1}$$

Where, $P_{i\&t}$ represents the joint output probability of the image modality and text modality; $C(\cdot)$ represents the output probability after the extraction of features by the convolutional neural network; $R(\cdot)$ represents the output probability after the extraction of features by the recurrent neural network ($Softmax(\cdot)$ was used for probability output in all the cases); L refers to the label of the image-text pair; N refers to the total number of disease categories.

2.5.2. Process of model application

In the process of model training, the convolutional neural network and recurrent neural network were used for feature extraction and probability output respectively, and in the training process, the weight and bias items of the network were updated whenever appropriate. In order to improve the credibility of the disease prediction results obtained by the model, knowledge database supervision was conducted on the basis of the training of the neural network. Path tracking for the knowledge database can visualize the final classification results, so as to perform model interpretation. In this study, the text modal information was used to link the knowledge graph. Before the linking process, Baidu's open source tool LAC was used to evaluate the importance of words, in order to find out notional words and discard non-notional words such as conjunctions.

The important notional words that were identified would be re-embedded into the network (using the word2vec training word vector). The obtained word vectors of important notional words were then used to link the knowledge graph. The linking process is shown in Eq (2).

$$d = O\left[W\left(T_{i\in t}^{text}\right), W\left(T_{j\in k}^{know}\right)\right] \tag{2}$$

Where, d represents the distance between the notional word vector $T_{i\in t}^{text}$ and the knowledge vector $T_{j\in k}^{know}$, $d \in [0,1]$; the distance measurement method $O[\cdot]$ adopts cosine similarity; t represents the notional word vector group; j represents the knowledge vector group. In this study, $d \geqslant 0.9$ indicates that the notional word is successfully linked with the knowledge graph.

After the text modality and the knowledge graph were successfully linked, the knowledge graph-assisted classification process was then carried out. In this process, the "image-text" multimodal identification model must be obtained first. The inference process of the disease classification model based on "image-text" multimodal collaborative representation and knowledge assistance is shown in Eq (3).

$$P_{i\&t\&k} = \sum_{i=0}^{N} M\left(T_i^{img} + T_i^{text}\right) + W\left(T_{j\to i}^{know}\right) \tag{3}$$

Where, $P_{i\&t\&k}$ represents the "image-text" multimodal joint output probability after combining with the knowledge graph; $M(\cdot)$ represents the "image-text" joint output; $W(\cdot)$ represents the initial probability after knowledge matching (the initial probability is $Softmax(n)$ in case of successful matching, where n represents the number of notional words that are successfully matched; otherwise the initial probability is $Softmax(0)$).

2.6. Experiment environment

Both the research experiment and control experiment were conducted in Ubuntu 18.04 environment (processor: Intel core i9 9820X; memory: 64G; graphics card: GeForce RTX 2080Ti 11G DDR6). The deep learning framework Pytorch, in combination with Cuda10.1, was used for training. In the experiment design and comparison process, the network batch-size of the training set and validation set was set to 16. The number of iterations of all network models was set to 50.

2.7. Measurement indicators

In this study, the various models were compared from 4 aspects, i.e., identification accuracy, identification precision, model sensitivity, and model specificity. The specific calculation equations are shown as follows (4–7).

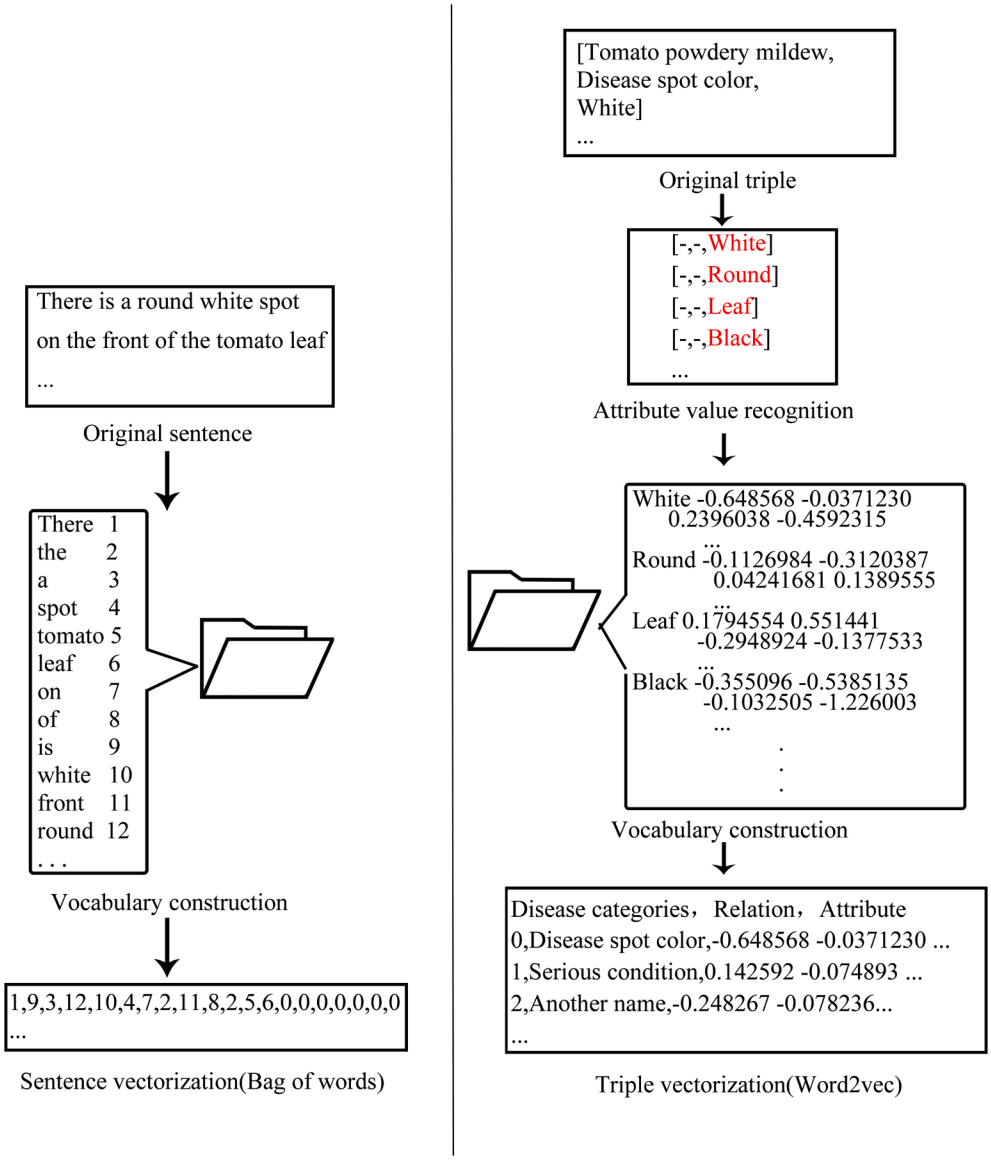

Fig. 2. Text vectorization process.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (5)$$

$$Sensitivity = \frac{TP}{TP + FN} \times 100\% \quad (6)$$

$$Specificity = \frac{TN}{FP + TN} \times 100\% \quad (7)$$

Where, TP refers to the number of individuals that are actually in category C and are correctly classified; FP refers to the number of individuals that do not belong to category C but are misclassified as category C; TN refers to the number of individuals that do not belong to category C and are correctly classified; FN refers to the number of individuals that belong to category C but are misclassified.

3. Results

3.1. Model comparison

The control networks for image modality in this study are AlexNet, VGG, ResNet, DenseNet, and MobileNet. The control networks for text modality are TextCNN, TextRNN, TextRNN_Att, and TextRCNN. In the process of "image-text" multimodal data joint training, different image modal and text modal network structures were combined randomly. The accuracy variation curve and Loss variation curve during the training process are shown in Fig. 4. The accuracy and loss are both the average of the six disease recognition results. The performance comparison on the classification results of control models is shown in Table 2. All the metrics in Table 2 are the average of test results six diseases in the test set.

In the model training process, DenseNet169 + TextRNN_Att showed the highest initial accuracy and also reached the highest accuracy after the training, and its Loss value dropped to the lowest level after the training. However, in the process of model application, due to the addition of the knowledge graph and the redistribution of the initial probability of diseases, ResNet50 + TextRNN_Att achieved the highest

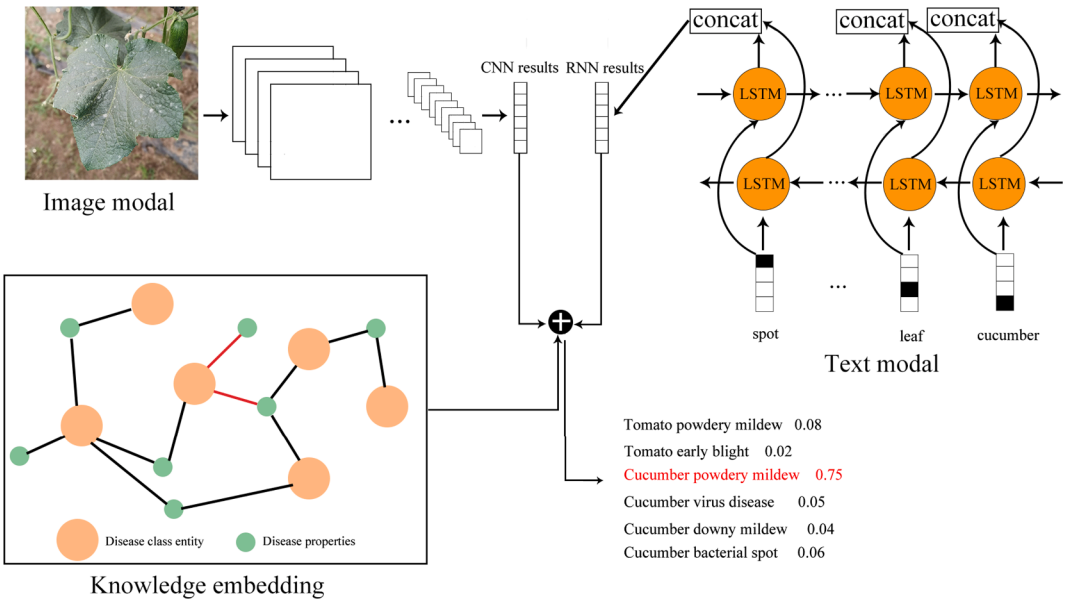

Fig. 3. The ITK-Net network structure diagram.

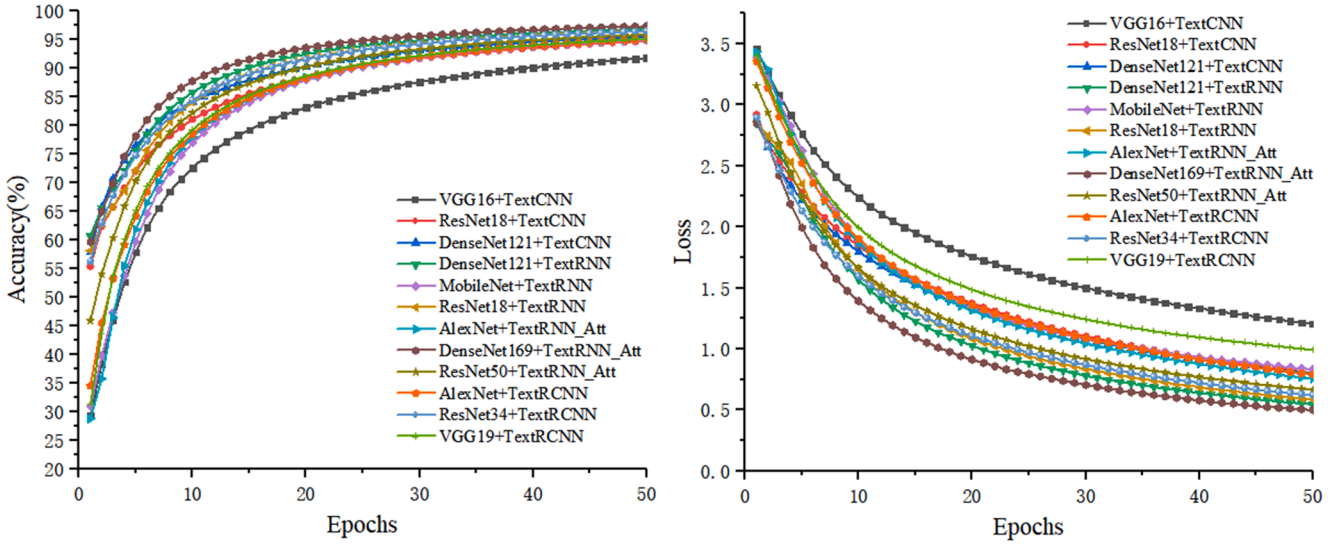

Fig. 4. The accuracy and Loss variation curves of control models.

accuracy, precision, sensitivity and specificity among all the control models; its confusion matrix of identification results is shown in Fig. 5. By analyzing the original image modal data and the original text modal data, it is found that DenseNet169 and ResNet50 can both accurately classify the images with large differences in visual information, but among two or several groups of images with insignificant differences in visual information (such as cucumber downy mildew and cucumber bacterial spot), the identification probability is very close and not accurate enough. In general, compared to DenseNet169, ResNet50 has a larger difference in identification probability for the disease images with similar visual information, and therefore, it can effectively distinguish similar diseases with the aid of knowledge graph. In order to minimize the potential impact of the small test set size on the recognition results, we re-divided the training set and the test set according to a 1:1 ratio, and then conducted a statistical test by means of 5 × 2 fold cross-validation experiment on the optimal combination of ResNet50 + TextRNN_Att. The test results are shown in Table 3. The results show that the proposed method is not sensitive to data segmentation, and have good stability and high identification accuracy.

3.2. Visualization of model inference process

3.2.1. Visualization of image modality inference

The disease images provide features for the model in the form of dense vectors, and the feature extraction part of the classification model is similar to a filter. The regions of significant interest in the visualized feature extraction layer is helpful for understanding the features learned by the model, so as to further analyze the disease features of the images. In this section, Grad-cam (Selvaraju et al., 2017) and Grad-cam++ (Chattopadhay, 2018) were used respectively to visualize the location of the feature interest points of the final model, as shown in Fig. 6.

In all the control models, AlexNet, ResNet18 and DenseNet121 were combined with different text modality models respectively to create the final models. Since the image modal and text modal parameters were updated concurrently during the training process, even if the image modality used the same network structure, the final model focus would

Table 2
Model performance comparison.

Models	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)
VGG16 + TextCNN	98.72	96.64	96.16	99.22
ResNet18 + TextCNN	99.45	98.45	98.52	99.67
DenseNet121 + TextCNN	98.53	95.80	95.42	99.11
DenseNet121 + TextRNN	98.72	96.49	96.42	99.22
MobileNet + TextRNN	97.62	94.41	93.45	98.53
ResNet18 + TextRNN	99.08	97.74	97.69	99.44
AlexNet + TextRNN_Att	98.53	95.74	96.02	99.12
DenseNet169 + TextRNN_Att	99.27	97.97	97.82	99.56
ResNet50 + TextRNN_Att	**99.63**	**99.00**	**99.07**	**99.78**
AlexNet + TextRCNN	97.99	94.50	93.60	98.78
ResNet34 + TextRCNN	98.72	96.56	96.25	99.22
VGG19 + TextRCNN	98.90	96.96	96.93	99.34

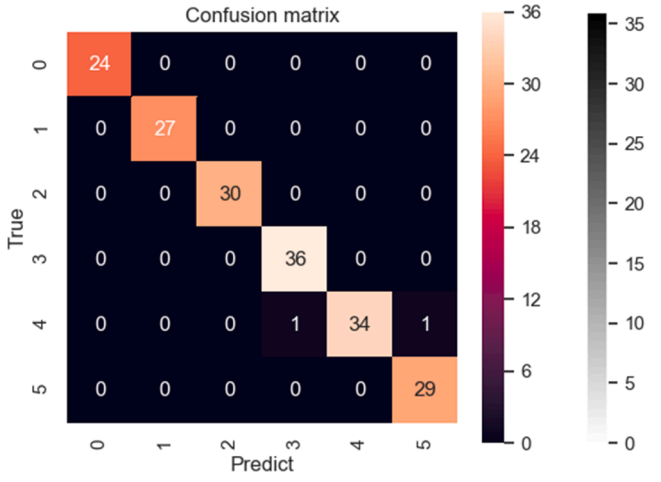

Fig. 5. Identification results of ResNet50 + TextRNN_Att. Note: 0 represents tomato powdery mildew; 1 represents tomato early blight; 2 represents cucumber powdery mildew; 3 represents cucumber virus disease; 4 represents cucumber downy mildew; 5 represents cucumber bacterial spot.

not be the same. The model focus of MobileNet has a relatively larger scope with a weaker image gradient, so its ultimate identification effect is the worst. Although VGG16 + TextCNN has the highest number of matched model focus points, its ultimate identification effect is still not good due to the occurrence of non-diseased areas. ResNet50 + TextRNN_Att has the highest degree of focus matching and the highest precision, so its ultimate identification effect is the best. In summary, all the image modality models, when used alone, can identify the real disease areas and achieve accurate identification based on the disease areas.

3.2.2. *Visualization of text modality inference*

The text modality provides features for the model in the form of sparse vectors, and its components are all natural language descriptions. In order to highlight the important features in the process of model inference, the elimination method was used in this section to visualize the inference process, that is, to eliminate vectorized texts one by one so as to infer which disease category the remaining text belongs to in accordance with the change of the result. A total of 4 models were selected: VGG16 + TextCNN, ResNet34 + TextRCNN, ResNet101 + TextRNN and ResNet50 + TextRNN_Att. These 4 models cover all the text modality models in the control experiment. The two disease categories with the highest error rate in the final prediction results (cucumber downy mildew and cucumber bacterial spot) were selected for inference visualization. The specific inference results are shown in Table 4.

In Table 4, the words marked in bold type in the original sentences are the words with wrong final prediction results after inference using the elimination method, and the other words refer to those with no error after inference. After analyzing the error category, it is found that the sentence about cucumber downy mildew was judged as early tomato blight after eliminating the word “cucumber”, was judged as cucumber early blight and cucumber bacterial spot after eliminating the word “yellow-green”, and was judged as cucumber virus disease after eliminating the word “light brown”, indicating that the corresponding words are all attributes of high importance among the real disease attributes. The sentence about cucumber bacterial spot disease was judged as tomato early blight after eliminating the word “cucumber”, and was judged as cucumber downy mildew after eliminating the word “yellow” or “round”, indicating that the corresponding words are also of high importance. Therefore, the method described in this section can not only validate the performance of individual models in the text modality but also map the important disease features described in natural language with the image modal features, thereby explaining the basis for the judgment of the disease category.

3.2.3. *Visualization of knowledge graph inference*

The knowledge graph provides a basis for the final judgment of the model. In the inference process, the text modality first makes a judgment by extracting the text features in the neural network and then maps the text modality with knowledge graph after evaluating the importance of words. This section presents the mapping results of part of the original texts with the knowledge graph disease attributes for ResNet50 + TextRNN_Att, as shown in Table 5.

It can be seen from Table 5 that after adding the knowledge graph into the “image-text” multimodal identification model, the disease knowledge involved in the original description text can be extracted. By analyzing the extracted disease attribute knowledge, it is found that the disease attribute is indeed a key attribute of the corresponding disease category, which confirms the effectiveness of the knowledge graph. Although there were also invalid triple matches, the overall matching success rate can basically meet the functional requirements. Since this study redefined the probability of the disease category based on the extracted disease knowledge triples, the corresponding operation could interfere with the identification results obtained by the training of the neural network so as to eventually influence the final identification results.

4. Discussions

The disease identification model based on the “image-text” multimodal collaborative representation and knowledge assistance (ITK-Net), which fuses image modal features with text modal features and incorporates the knowledge graph for supplementation, can fully utilize the various disease features so as to eventually realize the integration of

Table 3
Cross-validation result of ResNet50 + TextRNN_Att.

Dataset No.	1	2	3	4	5	6	7	8	9	10
Accuracy(%)	99.18	99.37	99.33	99.22	99.33	99.45	99.37	99.41	99.37	99.41

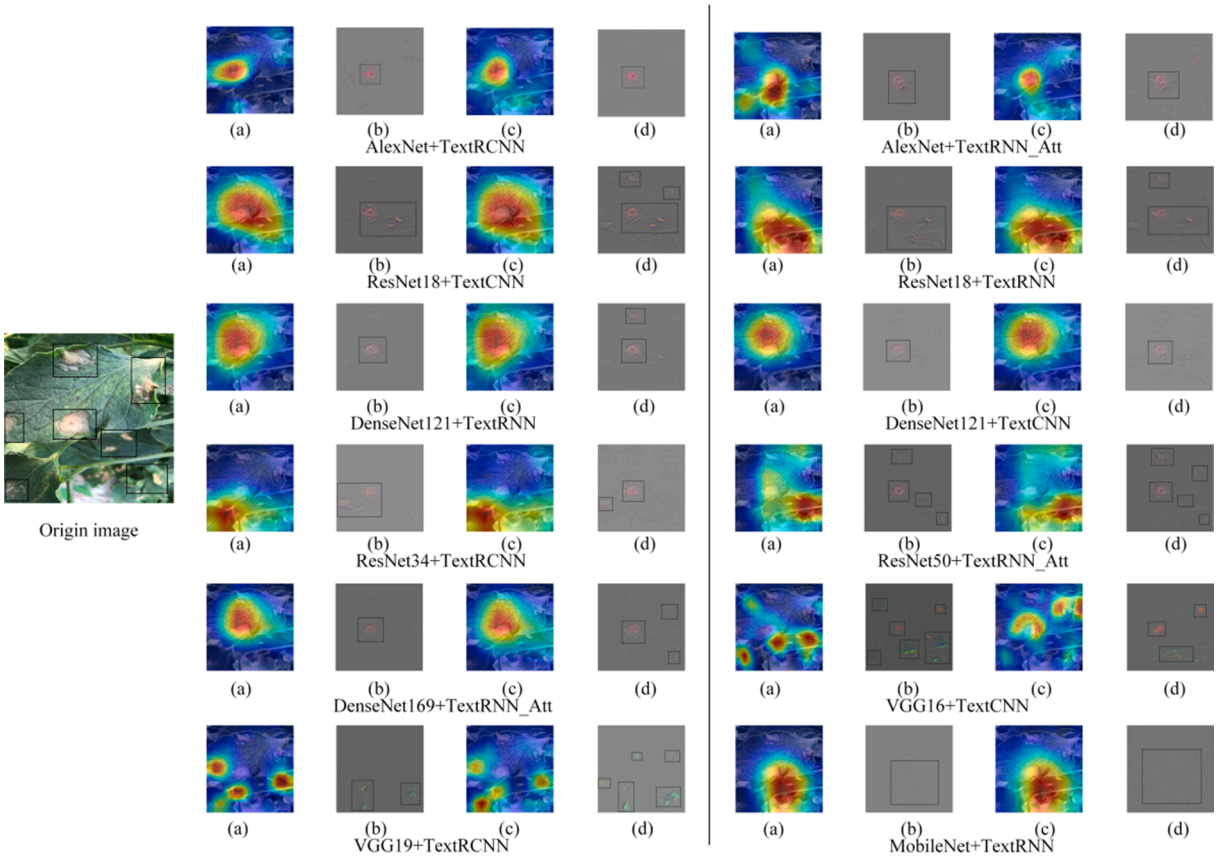

Fig. 6. Image modality visualization. Note: the locations of disease spots are manually labeled in the original images: (a) represents the visualization result of Grad-cam; (b) represents the image gradient visualization result of Grad-cam; (c) represents the visualization result of Grad-cam++; (d) represents the image gradient visualization result of Grad-cam++.

Table 4
Visualization of text modality inference.

Models	Text of cucumber downy mildew	Text of cucumber bacterial spot
VGG16 + TextCNN	On the adaxial surface of **cucumber** leaves there are **yellow-green** square patches with **light brown** patches on the edge	There are a few **yellow round** spots on the top of cucumber leaves
ResNet34 + TextRCNN	On the adaxial surface of **cucumber** leaves there are yellow-green square patches with light brown patches on the edge	There are a few yellow round spots on the top of cucumber leaves
ResNet101 + TextRNN	On the adaxial surface of cucumber leaves there are yellow-green square patches with light brown patches on the edge	There are a few **yellow round** spots on the top of cucumber leaves
ResNet50 + TextRNN_Att	On the adaxial surface of cucumber leaves there are **yellow-green** square patches with light brown patches on the edge	There are a few yellow round spots on the top of **cucumber** leaves

images, text and knowledge for disease identification. In this paper, inference visualization was performed for the three modalities. In the image modality, the model is expected to accurately locate and learn the disease features as far as possible. In the text modality, as the disease features were described in human natural language, they might be subjected to omissions or errors. The future research should focus on the preprocessing of natural language description, such as normalizing the description text and expanding single features to overall features. In the knowledge modality, as the disease knowledge was collected from Baidu Encyclopedia and other related websites, the description language was highly refined. Thus, there might be the problem of low matching success rate. In this study, a knowledge graph consisting of 6 tomato and cucumber diseases was established, and the future research should consider further optimizing the use of the knowledge graph and improving its usability.

Compared with a single image input, after adding text modal data, the accuracy of model recognition is improved. The disease knowledge graph enhances the interpretability of the model recognition results. However, text description increases the time for sample labeling. The generation of the knowledge graph requires professional plant protection knowledge. These works undoubtedly limit the ease of use of the proposed method. Future work could focus on how to eliminate these limitations, such as research on automatic generation tools for disease texts and the establishment of a public disease domain knowledge graph.

5. Conclusions

The disease image modality alone usually cannot cover all the effective information needed for accurate identification of the disease category. The fusion of features realizes information supplementation by gathering information from two modalities, i.e., image modality and text modality, and is therefore helpful for improving the precision of disease identification results and the robustness of the disease identification model. The knowledge vector redistributes the initial probability of disease categories by text matching, which can help improve the identification accuracy of the model and perform model interpretation. In this study, a knowledge graph consisting of 6 tomato and cucumber diseases was established. Among the various disease identification models based on “image-text” multimodal collaborative representation

Table 5
Visualization of knowledge graph inference.

Disease category	Original sentence	Mapping result
Tomato powdery mildew	There are white powdery spots on the front center of tomato leaves.	[Tomato powdery mildew, disease spot color, white] [Cucumber powdery mildew, disease spot color, white] [Cucumber bacterial spot, disease spot color, withered white]
Tomato early blight	There are a large number of bright brown ring spots on the upper part of the back side of tomato leaves.	[Tomato early blight, disease spot color, brown] [Cucumber bacterial spot, disease spot color, yellow–brown] [Tomato early blight, alias, ring spot disease] [Tomato early blight, disease spot shape, ring spot]
Cucumber powdery mildew	There are small light-white spots on the upper part of the front side of cucumber leaves and white spots on the lower part of the leaves.	[Cucumber virus disease, infectious virus, cucumber mosaic virus] [Cucumber powdery mildew, disease spot color, white] [Cucumber powdery mildew, disease spot color, white] [Cucumber bacterial spot, disease spot color, withered white]
Cucumber virus disease	The front side of cucumber leaves is yellow and green, and the leaves are wilted.	[Cucumber virus disease, infectious virus, cucumber mosaic virus] [Cucumber downy mildew, disease spot color, yellow]
Cucumber downy mildew	There are a large number of regular rectangular spots on the front side of cucumber leaves. The color of disease spots is yellow-green.	[Cucumber virus disease, infectious virus, cucumber mosaic virus] [Cucumber downy mildew, disease spot color, yellow-green] [Cucumber bacterial spot, disease spot color, yellow-green]
Cucumber bacterial spot	There are yellow disease spots on the front side of cucumber leaves, and the center of the disease spot withers.	[Cucumber virus disease, infectious virus, cucumber mosaic virus] [Cucumber downy mildew, disease spot color, yellow] [Tomato powdery mildew, leaf, early withering] [Cucumber powdery mildew, leaf, withered] [Cucumber bacterial spot, late stage of disease spot, withered and perforated]

and knowledge assistance (ITK-Net), the highest identification accuracy, precision, sensitivity, and specificity are 99.63%, 99%, 99.07%, and 99.78% respectively. In order to fully illustrate the effectiveness of the model, semantic interpretation was conducted from three aspects, i.e., image modality, text modality, and knowledge graph. In general, this research provides a method for disease identification by utilizing multimodality and provides semantic explanatory interpretation for the identification mechanism of the model.

CRediT authorship contribution statement

Ji Zhou: Writing - original draft. **Jiuxi Li:** Visualization, Investigation. **Chunshan Wang:** Writing - review & editing, Supervision. **Huarui Wu:** Data curation. **Chunjiang Zhao:** Methodology. **Guifa Teng:** Validation, Software.

Declaration of Competing 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.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFD1100602, and in part by the Beijing Municipal Science and Technology Project under Grant Z191100004019007, and in part by China Agriculture Research System of MOF and MARA under Grant CARS-23-C06, and in part by the Hebei Province Key Research and Development Project under Grant 20327402D,19227210D, and in part by the National Natural Science Foundation of China under Grant 61871041.

References

Frome, G.S., Corrado, J., Shlens, S., Bengio, J.D., Mikolov, T., 2013. Devise: A deep visual-semantic embedding model. NIPS 2, 3.
Akata, Z., Reed, S., Walter, D., et al., 2015:. Evaluation of output embeddings for fine-grained image classification. Proceedings of the IEEE conference on computer vision and pattern recognition. 2927–2936.
Chattopadhay A, Sarkar A, Howlader P, et al. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks[C]//2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018: 839-847.
Dai, H., 2020. What do CNN neurons learn: Visualization & Clustering. arXiv preprint arXiv:2010.11725.
Erhan, D., Bengio, Y., Courville, A., et al., 2009. Visualizing higher-layer features of a deep network. University of Montreal 1341 (3), 1–13.
Zhong, F., Chen, Z., Zhang, Y., Xia, F., 2020. Zero- and few-shot learning for diseases recognition of Citrus aurantium L. using conditional adversarial autoencoders. Comput. Electron. Agric. 179, 105828. https://doi.org/10.1016/j.compag.2020.105828.
Ghosal S, Blystone D, Singh A K, et al. An explainable deep machine vision framework for plant stress phenotyping. Proceedings of the National Academy of Sciences, 2018, 115(18): 4613-4618.
He, X., Peng, Y., 2017:. Fine-grained image classification via combining vision and language[C]//Proceedings of the IEEE. Conference on Computer Vision and Pattern Recognition. 5994–6002.
He, X., Peng, Y., 2020. Fine-grained visual-textual representation learning. IEEE Trans. Circuits Syst. Video Technol. 30 (2), 520–531.
Tobias, J., Springenberg, Alexey Dosovitskiy, Brox, et al., 2014. Striving for simplicity. The All Convolutional NET. eprint arxiv.
Ferentinos, Konstantinos P., 2018. Deep learning models for plant disease detection and diagnosis. Comput. Electron. Agric. 145, 311–318.
Lai Q, Khan S, Nie Y, et al. Understanding More about Human and Machine Attention in Deep Neural Networks. arXiv e-prints, 2019: arXiv: 1906.08764.
Ma, Juncheng, Du, Keming, Zheng, Feixiang, Zhang, Lingxian, Gong, Zhihong, Sun, Zhongfu, 2018. A recognition method for cucumber diseases using leaf symptom images based on deep convolutional neural network. Comput. Electron. Agric. 154, 18–24.
Picon, Artzai, Alvarez-Gila, Aitor, Seitz, Maximiliam, Ortiz-Barredo, Amaia, Echazarra, Jone, Johannes, Alexander, 2019. Deep convolutional neural networks for mobile capture device-based crop disease classification in the wild. Comput. Electron. Agric. 161, 280–290.
Smilkov D, Thorat N, Kim B, et al. Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825, 2017.
Simonyan, K., Vedaldi, A., Zisserman, A., 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. Computer ence.
Selvaraju, R.R., Cogswell, M., Das, A., et al., 2017:. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE international conference on computer vision. 618–626.
Savary, Serge, Willocquet, Laetitia, Pethybridge, Sarah Jane, Esker, Paul, McRoberts, Neil, Nelson, Andy, 2019. The global burden of pathogens and pests on major food crops. Nat. Ecol. Evol. 3 (3), 430–439.
Xu, S., Venugopalan, S., Sundararajan, M., 2020:. Attribution in scale and space. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 9680–9689.
Xu, H., Qi, G., Li, J., et al., 2018. Fine-grained Image Classification by Visual-Semantic Embedding. Twenty-Seventh International Joint Conference on Artificial Intelligence IJCAI-18.