Research on Crop Disease Identification Methods Based on Multimodal Data Fusion and CLIP Model

Di Lu[1], Tongjin Liu[1], Tengfei Zhang[1], Shengzhi Liu[1], Xiyu Zhang[1], Jiahao Lei[1], Fanghua Chen[1]

[1]Xinjiang College of Science & Technology, Korla 841000, Xinjiang, China

Corresponding author: Tongjin Liu (e-mail: 13779662707@163.com)

ABSTRACT Crop disease is a major threat to food security, while current identification methods often suffer from low accuracy and weak generalization under complex field conditions. This paper proposes a multimodal crop disease identification model integrating image data and textual symptom descriptions. First, a multimodal dataset covering 12 common crop diseases is constructed, including 8,000 high-quality disease images and 8,000 symptom texts. Second, a Contrastive Language-Image Pre-training model is used to extract cross-modal features, and an attention mechanism is introduced to achieve deep fusion of image and text representations. Third, a multi-scale feature pyramid network is designed to capture disease characteristics at different granularity levels, while a focal loss function is introduced to address class imbalance. Experimental results show that the proposed method achieves 96.8% accuracy in the 12-category disease identification task, which is 8.3 percentage points higher than the single-modal method. The recall and F1 score reach 95.7% and 96.2%, respectively, and the inference speed reaches 42 frames per second. Ablation experiments show that the multimodal fusion module contributes a 4.5% performance improvement.

INDEX TERMS crop disease identification, multimodal data fusion, clip model, attention mechanism, smart agriculture

I. INTRODUCTION

THE continued growth of the global population places higher demands on food production, yet pests and diseases cause an average annual crop yield loss of over 20%, directly undermining food security and the high-quality supply of raw materials for manufacturing excellence. Integrating advanced fiber morphology analysis into agricultural monitoring can enhance early disease detection, ensuring both sustainable agricultural development and the structural integrity of industrial fiber resources. If diseases such as wheat rust, rice sheath blight, and corn leaf spot are not detected and controlled in time, they can lead to reduced yields or even total crop failure in some areas. Early symptoms of diseases are not obvious, and manual inspection relies on expert experience, which is costly and inefficient, and difficult to cover large-scale planting areas [1,2]. Computer vision technology provides a solution for automatic disease identification, but existing methods mainly rely on single image data, ignoring the textual description knowledge accumulated by plant protection experts. The accuracy of identification drops significantly in complex scenarios such as changes in light, background interference, and disease similarity, and the generalization ability is limited.

Convolutional neural networks are deep learning models that have led to advancements in property tasks of image classification but the multidimensional properties of diseases cannot be fully captured using the visual features. It is not only that disease diagnosis depends on a visual input like the color and texture of the leaf, but it must incorporate a semantic input like the location of the disease, the development of symptoms, and morphology of the lesions. Multimodal learning improves the representational capability of the model by combining the data of varied sources. The CLIP model relies on a contrastive learning paradigm to complete a unified embedding space mapping between text and images to offer a simple framework of cross-modal feature extraction [3,4]. Nevertheless, the pre-training data that CLIP provides is primarily based on the general Internet cases and does not possess professional knowledge in the agricultural sphere of work, so it is challenging to transfer the knowledge of the domain in the direct application to disease recognition. The question of how to work out an efficient feature fusion strategy, which would help to take the most out of the complementary benefits of the visual and the textual modalities, has become one of the issues of

primary concern that must be addressed as soon as possible.

In this paper, we construct a multimodal dataset encompassing diverse crop diseases and propose a CLIP-based cross-modal feature extraction and attention-based fusion model to integrate agricultural pathological features with fiber morphology data. This approach enables a more comprehensive characterization of disease impact on raw material quality, thereby enhancing the precision of intelligent monitoring for manufacturing excellence. The model is based on the idea of deep semantic alignment between images and text, a bidirectional attention mechanism, gating units modulate contribution weights of various modalities, and a multi-scale pyramid network acquires the fine-grained feature representation of diseases. Focal loss function is added to address the issue of the imbalance of the training data in terms of classes and enhance the capacity of the model to detect low-frequency diseases [5]. The effectiveness of the proposed method is tested with experimental results, and ablation experiments determine the contribution to performance of each module to the technical support of intelligent disease diagnosis in agriculture.

II. RELATED WORK

Deep neural network-based crop disease identification systems are developed on the basis of the traditional machine learning. The earlier techniques used were based on manually constructed feature extraction operators, including color histograms, texture descriptors, and shape invariants, and used with support vector machines or random forest classifiers to obtain disease detection [6,7]. These techniques involve the involvement of the domain experts in feature engineering, they possess low generalization power, and sensitive to image quality. Convolutional neural networks have transformed the feature learning paradigm and are able to learn hierarchical feature representations bottom-up, automatically, as a result of end-to-end training. Scientists have used ImageNet models that have been pre-trained and fine-tuned to recognize diseases in agricultural settings. Mobile Lightweight network architecture Mobile network architecture is tailored to be lightweight in order to decrease the number of required computational resources without compromising recognition accuracy. This is made possible by the introduction of the attention mechanisms to direct the model to the lesion areas and eliminate the noise interference in the background. Though these techniques have been successfully used to produce good results on standard data, the weakness of a single visual modality limits the recognition performance in complex situations.

Multimodal learning is the way to integrate heterogeneous sources of data and increase the representational capability and strength of models. Computer image and text joint modeling have gained substantial ground in broad areas and contrastive learning strategies are attaining cross-modal correspondence by increasing the likeness of matches amongst sample pairs. CLIP model is trained with large scale image-text pairs to train the model to learn the general visual-language representation space, with strong zero-shot transfer abilities. The model has found extensive applications in methods like object detection, semantic segmentation, and image generation, yet its flexibility in vertical application is yet to be discovered [8,9]. Multimodal research within the agricultural domain is currently in its infancy. Across some works there are the attempts to integrate meteorological data, soil information, or sensor readings to predict disease, however, there are not many studies on deep fusion of text and images. Simple feature splicing approaches do not pay attention to the semantic connection between modalities and cannot make the most of complementary information. The attention mechanisms offer a solution to the cross-modal interaction through attainment of the transfer of information among the modalities through dynamically distributed weights. Most of the available attention modules, however, are one-way designs and do not model bi-directional semantic dependencies. Multiscale feature extraction has been found to be useful in object detection. This is because feature pyramid networks are grassrooted networks that includes multi-granular information through fusion of feature maps at varying levels. The use of the concept in the disease identification remains inadequate, particularly the absence of systematic studies on the collaborative design with multimodal fusion.

III. METHODS

A. CONSTRUCTION OF MULTIMODAL DISEASE DATASET

The paper gathered information about 12 prevalent diseases which comprise rust, powdery mildew, leaf spot and sheath blight of three staple crops namely wheat, corn and rice. Image acquisition was conducted in experimental fields within the provinces of Jiangsu, Henan, and Shandong. The pictures were captured with a Canon EOS 5D Mark IV camera at a resolution of 4096x2160 pixels to ensure that the underlying pathological features and fiber morphology were preserved without compression artifacts during the initial archival phase. At least 700 samples were taken in each disease which included the leaf morphological modification in the early stages of the disease, the middle stages of the disease, and the late stages of the disease as well as the images with a different lighting, as sunny, cloudy, and back lighting [10].

The text annotation was done in a joint work of five experts in the field of plant protection. The images are well accompanied by a structured description that includes four dimensions namely disease name, affected area, symptom characteristics and severity. The description of the symptoms is made in standardized terms, i.e., the irregular brown spots are observed on the edges of leaves, 2-5 mm in diameter, with dark brown periphery and grayish-white center, surrounded by yellow halo. To enhance data diversity and expand the initial 8,400 raw captures (approx. 700 per category), the original images underwent random cropping, 15-degree rotation, and 20% brightness adjustment. After

rigorous quality filtering, a balanced dataset of 8,000 augmented image-text pairs was finalized. To mitigate overfitting risks inherent in Transformer-based architectures with limited domain-specific data, we employed a combination of a 0.3 dropout rate, heavy weight decay (0.01), and the frozen backbone strategy for the CLIP encoders, ensuring the model generalizes well beyond the training samples.

B. CROSS-MODAL FEATURE EXTRACTION BASED ON CLIP

This paper uses the ViT-L/14 architecture as the feature extraction backbone network using the CLIP model. Although the CLIP image encoder downsizes the input to 336x336 pixels, the high-resolution originals allow for high-fidelity patch extraction and multi-scale feature reinforcement, ensuring that even after downsizing, the semantic integrity of the fine texture changes remains superior to images captured at lower native resolutions. Linear projection is used to project each patch into a 1024-dimensional vector. Position encoding makes use of a learnable parameter matrix that is appended to the patch embedding. The input is a visual branch that is 24 Transformer encoder layers with 16 heads in the multi-head self-attention mechanism. The encoder generates [CLS] as the global image representation that has a dimension of 1024.

For processing disease symptom descriptions, the text encoder utilizes a BPE tokenizer with a vocabulary of 49,152 tokens to translate the text into a sequence of tokens, a sequence size of which is constrained to 77. The tokens are mapped by a 512-dimensional word embedding layer and overlaid with positional encodings and fed to a 12-layer Transformer encoder with 8 attention heads. The text features are mined using the positions at the end of the sequence of the [EOS] tokens, and the dimension is set to 512. In the creation of cross-modal alignment, the image features are compressed into a 512 dimensions linear projection layer, and cosine similarity is computed between the image features and text features in a shared embedding space. The performance of various CLIP architectures is compared in disease features extraction as shown in Table 1.

The data in Table 1 show that the ViT-L/14@336px architecture achieves a similarity calculation accuracy of 96.2%. Although the feature extraction time is 118ms, it performs best in capturing fine-grained features of lesions compared to other architectures, and can effectively identify the texture changes and color gradient features of lesion edges.

C. ATTENTION MECHANISM-DRIVEN FEATURE FUSION STRATEGY

A cross-modal attention module is used to semantically match image and text features. This module has three sets of learnable weight matrices, which transform 512-dimensional image features into query representations and text features into key and value representations respectively. Attention weights are computed by taking the dot product of the query and key and then activated by a softmax activation function with a scaling factor to produce a distribution of weight that indicates the strength of a semantic association between image and text. These weights are imposed on the value representations, giving results of text-directed image feature enhancement which in turn are added to the original image features, using residual connections, without destroying underlying visual detail information.

Bidirectional attention determines the attention of the text to the image at the same time, thus producing an image-guided text feature description. The two directions make up the enhanced features that are joined together into a 1024-dimensional joint representation through concatenation before being fed into a gated fusion unit to adjust the adaptive weighting. The gated unit uses two-layer fully connected network with a dimension of the hidden layer of 256 and GELU activation. A sigmoid shaped output layer creates fusion weights ranging 0 to 1. These weights regulate the ratio of contribution of visual and semantic features and adapts the fusion strategy to various types of diseases dynamically. Diseases that are morphologically prominent, like leaf spot are assigned bigger weights in the visual branch and the text branch of rust which includes complex description of symptoms, increases its text branch influence. The concatenated feature vector is reduced to 512 dimensions and it makes the input representation to the classifier.

D. MULTI-SCALE FEATURE PYRAMID NETWORK

The system uses a pyramid structure of four layers to represent the features of the disease at various granularities. The last layer has a 3×3 convolutional kernel and a stride of 1 resulting in a 128 dimensional feature map with a receptive field of 7×7 pixels in size, detailing fine texture changes and details of color transition at lesion boundaries. The second layer applies a 5×5 convolutional kernel with 2-stride that yields a 256-dimensional feature map with a 15×15 pixel receptive field that extracts the entire morphological information of each lesion, such as circular, elliptical, or irregular outlines. The third layer is a 7×7 convolutional kernel with a stride of 4, which yields a 512-dimensional feature map with the receptive field of 31×31 pixels and determines the spatial patterns distribution and level of clustering of multiple lesions. The upper layer applies the global average pooling that will reduce the whole image into one 512-dimensional global semantic vector depicting the general severity of the disease.

At four scales, feature maps were upsampled and aligned. Crucially, the 512-dimensional text embedding is projected through a multi-layer perceptron (MLP) to match the channel dimensions of each pyramid level. These projected text features are then element-wise multiplied with the spatial feature maps at each scale. This cross-modal spatial injection allows the gating unit to dynamically amplify or suppress specific pixel-level features based on textual patho-

TABLE 1. Comparison of feature extraction performance of different CLIP architectures

Model Architecture	Image encoder layers	Text encoder layers	Feature Dimension	Extraction time (ms)	Similarity calculation accuracy (%)
ViT-B/32	12	12	512	28	89.3
ViT-B/16	12	12	512	45	91.7
ViT-L/14	24	12	1024→512	83	94.5
RN50x16	50	12	768	92	90.8
RN50x64	50	12	1024	156	92.1
ViT-L/14@336px	24	12	1024→512	118	96.2

logical descriptors, providing a structural basis for high-precision identification of diseases like rust that rely on fine-grained semantic cues. The final 1408-dimensional feature representation was fed into a three-layer fully connected classifier containing 512, 256 and 128 neurons in the hidden layers, and a dropout rate of 0.3 to avoid overfitting. The dataset was corrected with Focal Loss to deal with the imbalance in classes based on the focus factor of 2.0 and reduction factor of 0.25. The design assigned samples with a smaller incidence rate a higher weight gradient, such that the rust recall rate was higher in the test set of 93.6 as compared to 82.1 in the training set with powdery mildew covering 38 percent and rust covering 5 percent.

IV. RESULTS AND DISCUSSION

A. EXPERIMENTAL SETUP

The experiments were implemented on a server that had an NVIDIA RTX 4090 and 64GB RAM, Ubuntu 22.04 operating system, and PyTorch 2.0.1 was used as the deep learning package. The AdamW optimizer with initial learning rate of 0.0001, a weight decay factor of 0.01, and a batch size of 32 was used as model training. Learning rate schedule was done by cosine annealing, and the training cycle consisted of 100 epochs. Validation set was used to assess performance on a validation set after every 10 epochs and the best model weights were stored. Speed tests of inferences were conducted using one gpu and the statistical analysis of the frame rate of processing 1000 images was carried out.

B. MODEL PERFORMANCE COMPARISON ANALYSIS

At the testing stage, the test set was randomly picked to have 800 images, which would be divided equally among the 12 disease categories with an average of about 67 samples in each category. The testing environment had the same hardware set as the training environment and all background processes were closed so that there was no contention of resources. Once it loaded the pre-trained weights, the model was then set to evaluation mode, which disabled dropout and randomness of the batch normalization layer. Preprocessing of the input images was done in a standardized manner with the mean being [0.485, 0.456, 0.406] and standard deviation being [0.229, 0.224, 0.225]. In the inference stage, the batch size was set to 1. To facilitate real-world deployment where expert text is unavailable, the model operates in a 'prompt-based' inference mode. The text encoder utilizes predefined symptom templates (e.g., 'A leaf with [Disease Name] showing [Symptom]') to calculate similarity scores against the input image. This allows the model to leverage its cross-modal alignment capabilities to identify diseases based solely on visual input, matching it against the most likely textual pathological description learned during training. The predicted category, confidence score, and processing time for each image were recorded, and overall performance metrics were statistically analyzed.

The proposed method in Table 2 outperforms the large-scale single-modality baseline (ViT-L/14) by 6.6 percentage points with respect to its accuracy and the standard multi-modal ViT-BERT concatenation strategy by 5.6 percentage points. The recall rate is 95.7%, and this shows that it is good in detecting different diseases, especially rust which is limited in sample size. The F1 score is 96.2% which is an excellent precision recall balance. The inference rate has been slowed down to 42 frames a second but it is sufficient to achieve real time detection. It also decreases the speed by a maximum of 5 frames per second compared to simple attention fusion method and the time overhead caused by the performance improvement is acceptable.

C. ABLATION EXPERIMENTS AND CONTRIBUTION ASSESSMENT OF EACH MODULE

The efficacy of each component was confirmed through ablation experimentation by the systematic removal or substitution of crucial modules. The starting model utilized the ViT-L/14 architecture as a baseline image encoder and linear classifier, providing a high-capacity Transformer foundation before the addition of text features, cross-modal attention, bidirectional attention, gated fusion unit as well as multi-scale pyramid network in that order. All the configurations were trained with 100 epochs on the same training data, with the same hyperparameter setting and data augmentation strategies. Testing on the test set was done to test the performance metrics to make sure that there is fair assessment between experiments and controls.

Table 3 indicates that the accuracy of the introduction of text features increased by 1.7 percentage points (from 89.4% to 91.1%), which confirms the complementarity of multimodal data. The cross-modal attention system added a 2.4 percentage point benefit, which attained efficient alignment of images and text. Two-way attention also raised the accuracy further by 1.3 percentage points, which is the opposite guidance effect of text on pictures. The gated

TABLE 2. Performance Comparison of Different Methods in Disease Identification Tasks

Method	Accuracy (%)	Recall rate (%)	F1 score (%)	Inference speed (frames per second)
Feature splicing and fusion	91.2	89.8	90.5	51
Simple attention fusion	93.6	92.4	93.0	47
This article's method	96.8	95.7	96.2	42
ViT-L/14 (image-only baseline)	90.2	88.4	89.3	52
CLIP-ViT-L/14 (zero-shot)	84.5	82.7	83.6	48
Multimodal ViT + BERT (splicing)	91.2	89.8	90.5	51

TABLE 3. Performance Contribution Analysis of Each Module in the Ablation Experiment

Model Configuration	Text features	Cross-modal attention	Bidirectional attention	Gating fusion	Multiscale pyramids	Accuracy (%)
Baseline model	×	×	×	×	×	89.4
+Text Features	✓	×	×	×	×	91.1
1. Cross-modal attention	✓	✓	×	×	×	93.5
Bidirectional attention	✓	✓	✓	×	×	94.8
+Gating Fusion	✓	✓	✓	✓	×	95.6
Complete model	✓	✓	✓	✓	✓	96.8

fusion unit was able to improve the results by 0.8 percentage point, adjusting the weights of various modalities dynamically. The greatest contribution was on the multi-scale pyramid network, which increased the accuracy by 1.2 percentage points; multi-granularity feature extraction gave the model the capability to identify lesions of varying sizes.

V. CONCLUSION

This paper discusses the limitations of unimodal data representation in crop disease detection and its lack of flexibility in complex field conditions, which often fails to capture the intricate fiber morphology variations essential for manufacturing excellence. By integrating multimodal data, the proposed framework enhances the characterization of pathological impacts on raw material quality, ensuring more robust agricultural and industrial performance. It suggests a multimodal identification framework, which amalgamates visual images and text descriptions. The CLIP model will be applied to extract cross-modal features by creating a high quality dataset with different types of crop diseases. A feature fusion strategy that is based on attention and a pyramid network of different scales is combined to obtain the disease feature representation on various granularities. The results of the experiments prove that the proposed approach has excellent performance regarding the ability to identify and generalize well. Experiments of ablation prove the usefulness and necessity of every module. This study offers a viable technical route on early diagnosis of diseases in smart agriculture, but has limitations. The model must enhance its detection of rare diseases by integrating expert-annotated fiber morphology data, ensuring the preservation of raw material quality for manufacturing excellence. Furthermore, the inference speed should be optimized to facilitate real-time deployment on edge devices within complex agricultural and industrial environments.

AUTHOR CONTRIBUTIONS

Conceptualization –Di Lu, Tongjin Liu, Tengfei Zhang, Shengzhi Liu, Xiyu Zhang, Jiahao Lei and Fanghua Chen; methodology – Di Lu, Tongjin Liu, Shengzhi Liu, Xiyu Zhang, Jiahao Lei and Fanghua Chen; investigation – Di Lu, Tongjin Liu, Tengfei Zhang, Shengzhi Liu, Xiyu Zhang, Jiahao Lei and Fanghua Chen; writing-original draft preparation – Di Lu, Tongjin Liu, Tengfei Zhang, Shengzhi Liu, Xiyu Zhang, Jiahao Lei and Fanghua Chen. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] S. Yang, Q. Feng, W. Yan, et al., "Small sample crop disease identification using multimodal guided visual Transformer, " Transactions of the Chinese Society of Agricultural Engineering, vol. 41, no. 6, p. 195, 2025, doi: 10.11975/j.issn.1002-6819.202409189.

[2] Z. Wang, F. Ma, Y. Zhang, et al., "Crop disease identification based on attention mechanism and multi-scale lightweight network, " Transactions of the Chinese Society of Agricultural Engineering, vol. 38, no. S01, pp. 176-183, 2022, doi: 10.11975/j.issn.1002-6819.2022.z.020.

[3] S. Deng, J. Zhu, Y. Hu, and et al.Tomato Leaf Disease Identification Framework FCMNet Based on Multimodal Fusion, "Plants (2223-7747), " 2025; 14(15), doi: 10.3390/plants14152329.

[4] H. Dong, "Crop disease identification based on transfer learning and residual network, " Computer Science and Application, vol. 11, no. 04, pp. 1165-1172, 2021, doi: 10.12677/csa.2021.114120.

[5] C. Yang, Y. Zhang, M. Zhao, et al., "Research progress on early crop disease detection and identification technology based on infrared thermography, " Laser Journal, vol. 41, no. 6, p. 4, 2020, doi: 10.14016/j.cnki.jgzz.2020.06.001.

[6] V. Jaiswal, V. Ranjan, S. Gupta, et al., "Field Management and Crop Disease Identification System, " International Journal of Engineering Research and, 2020, doi: 10.17577/ijertv9is060165.

[7] H. Yevlekar, P. Deore, P. Patil, et al., “Smart and Integrated Crop Disease Identification System, ” International journal of scientific research in science, engineering and technology, vol. 5, pp. 189-193, 2020, doi: 10.32628/IJSRSET2051040.
[8] D. Ahmed, “Crop Disease Identification Using Ensemble Deep Learning, ” International Journal of Science, Engineering and Technology, vol. 12, no. 6, pp. 1-10, 2024, doi: 10.61463/ijset.vol.12.issue6.342.
[9] C. Wang, Y. Xia, L. Xia, et al., “Dual discriminator GAN-based synthetic crop disease image generation for precise crop disease identification, ” Plant Methods, vol. 21, no. 1, pp. 1-22, 2025, doi: 10.1186/s13007-025-01361-0.
[10] L. Hao, Q. Weigen, and Z. Lichen, “Improved ShuffleNet V2 for Lightweight Crop Disease Identification, ” Journal of Computer Engineering & Applications, pp. 58(12), 2022, doi: 10.3778/j.issn.1002-8331.2111-0457.