ENGINEERING Agriculture

Engineering Journals

RESEARCH ARTICLE

2027 / Volume 14 / Issue 2 / 27728

A novel detection method based on a self-supervised framework integrating multi-scale spatial relationships for tomato leaf disease

Junyan ZHU[1], Enqi YANG[1], Chunji SHI[1], Changcai YANG[1], Liangquan WU[2], Yaohai LIN (✉)[1]

1 College of Computer and Information Sciences, Fujian Agriculture and Forestry University, Fuzhou 350002, China.
2 College of Resources and Environment, Fujian Agriculture and Forestry University, Fuzhou 350002, China.

KEYWORDS

Tomato leaf disease detection, self-supervised learning, crop disease recognition, mask autoencoder

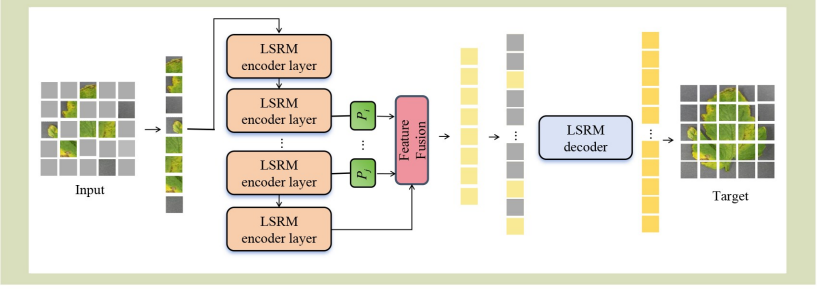
HIGHLIGHTS

- A novel method is proposed for tomato leaf disease detection.
- The proposed model integrates cross-layer feature fusion mechanism can improve model's ability to capture semantic details of subtle disease features.
- The proposed model doesn't rely on large annotated images and outperforms the baseline.
- The proposed model captures complex structural information between different regions in image.

Received January 19, 2026;
Accepted June 10, 2026.

Correspondence: linyaohai@fafu.edu.cn

GRAPHICAL ABSTRACT



ABSTRACT

Accurate diagnosis of tomato leaf diseases is essential to ensure optimised tomato production and reduce losses. However, their fine-grained visual features are challenging to capture and are susceptible to interference from complex agricultural environments. Conventional object detection methods rely on large-scale annotated datasets, which limit detection efficiency and accuracy. This study proposes a novel method, the Cross-layer Spatial Feature Fusion Mask Autoencoder (CSF-MAE), based on a self-supervised framework that integrates multi-scale spatial relationships to precisely detect nine distinct types of tomato leaf disease. The proposed method does not require many annotated samples. It exhaustively captures local and global image features using learnable spatial relationship modules and cross-layer feature fusion mechanisms, thereby enhancing fine-grained detection performance and environmental adaptability. Comparative experiments demonstrated that the proposed model outperformed the baseline, achieving 99.05% accuracy, 99.01% precision, 99% recall and 99% *F*1-score. The findings revealed that this technique displayed outstanding robustness across intricate backgrounds and high-occlusion scenes, efficiently accomplishing accurate detection of early tomato leaf diseases, with significant practical application value.

© The Author(s) 2027. Published by Higher Education Press. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0)

1 Introduction

Crop diseases pose a substantial risk not only to regional food security but also to the stability of the global food supply chain[1]. As one of the most widely cultivated crops, the tomato provides a rich source of nutrition for humans and has considerable economic value[2–4]. However, due to various factors, such as environmental changes, tomato leaves are prone to disease, which affects yield and quality. Traditional identification of tomato leaf diseases primarily relies on the expertise of farmers and specialists; however, such manual methods are often inefficient, labour-intensive and liable to error[5]. These restrictions are particularly pronounced in large-scale agricultural monitoring, where delayed interventions can exacerbate disease spread. Consequently, the rapid and accurate determination of tomato leaf diseases is crucial for implementing timely and effective control measures.

In recent years, with the rapid advancement of artificial intelligence, deep learning methods have catalysed extensive research into their applications for crop disease recognition. Deep learning–based image analysis has emerged as a powerful tool, enabling rapid, precise and non-invasive disease diagnosis. By leveraging deep learning's superior pattern recognition capabilities, it is possible to upgrade the efficiency and accuracy of disease diagnosis while minimising misclassification rates, thereby offering intelligent solutions for sustainable agricultural production and food security. Deep learning methods have been widely applied in the identification of crop pests and diseases, such as wheat[6–8], maize[9–11], rice[12–14] and apples[15–17].

Deep learning methods have also achieved breakthroughs in detecting tomato leaf diseases, especially object detection methods such as the YOLO series. For instance, Wang et al.[2] improved upon YOLOv10n by introducing a deformable convolutional network (DCN) to the YOLOv10n backbone network and incorporating a bidirectional feature pyramid network (BiFPN) on top of the FPN + PAN structure, identifying four types of tomato leaf diseases with precision increased from 85.1% to 87.2%, recall from 86.3% to 89.1%, and mean average precision (mAP) from 87.4% to 91.3% compared to the baseline model. Sun et al.[18] introduced the CSWinTransformer (CSWinT) module into the YOLOv8n network architecture, achieving a mean average precision (mAP50) of 97.2% for tomato leaf disease detection. To address challenges such as large model sizes and slow detection speeds encountered in practical applications, He and Tong[19] incorporated a novel, efficient multi-scale attention (EMA) mechanism into the deeper layers of the YOLOv8n architecture. They augmented the C2f module into a RepViT Block (RVB) using decoupled tokens and channels. Experimental findings demonstrated that the proposed model decreased the number of parameters by 43% and the computational load by 50%. Liu et al.[20] expanded YOLOv8s by implementing Mamba2 into the neck network for tomato leaf disease recognition, and yielded improvements of 3.0%, 4.8% and 4.3% in *F*1-score, mAP@0.5 and mAP@0.5:0.95, respectively, compared to YOLOv8s. Wang et al.[21] introduced the multi-path inverted residual unit (MPIRU) and the dynamic focusing attention framework (DFAF) into the YOLOv10 architecture to overcome the drawbacks of existing detection approaches in terms of accuracy, computational efficiency and scalability. Yang et al.[22] designed the MWD–YOLO11n network based on a multi-scale dynamic attention (MSDA) mechanism, wavelet convolution (WTConv) and a dynamic sampling operator (DySample). The MWD–YOLO11n network can detect seven types of tomato leaf diseases using 6370 tomato leaf images, achieving 89.3%, 83.2%, 88.8% and 77.9% in accuracy, recall, mAP50 and mAP50-95, respectively. Ye et al.[23] proposed a refined multidimensional tomato disease identification model based on YOLOv8 to enhance the accuracy of tomato disease identification. Lee et al.[24] improved YOLOv5m by using C3NN modules and a BiFPN architecture. Compared with YOLOv5m, it achieved improvements in precision by 1.36%, recall by 4.99%, *F*1-score by 3.18%, mAP.5 by 1.74% and mAP.5:.95 by 3.26%. Liu et al.[25] utilised an improved YOLOv8 network that integrated the multi-head self-attention (MHSA) and shuffle attention (SA) modules to increase the accuracy of tomato leaf disease detection.

The YOLO series achieves an excellent balance between speed, accuracy and complexity. Similarly, the aforementioned object detection methods have made significant progress in crop disease classification. Despite the quantum leap made by supervised deep learning methods such as the YOLO series in increasing the crop disease recognition accuracy, they remain constrained by the inherent flaws of the network. The YOLO series faces the following shortcomings:

(1) Difficulty in detecting small objects. Due to potential loss of details in deep network features, the YOLO series performs poorly at detecting small, blurry or occluded objects, often leading to missed detections.

(2) Robustness issue in convoluted scenarios. When faced with drastic lighting variations, cluttered backgrounds, adverse weather conditions or adversarial attacks, the model's performance may degrade substantially.

(3) Dependence on training data. High-performance YOLO models typically require large-scale, high-quality and accurately labelled datasets for training, and the cost of acquiring and labelling such data is high.

The aforementioned shortcomings of the YOLO series, especially its reliance on annotated datasets, have somewhat limited the model's generalisation capability, as acquiring high-quality annotated datasets is not only expensive but also challenging due to significant distribution shifts across regions and crop types[26–28]. In contrast, self-supervised learning (SSL) alleviates dependence on annotated data by designing pretext tasks that enable models to learn inherent patterns and structures directly from the data itself[29–31]. With the development of visual SSL, particularly masked image modelling (MIM), and its expanding applications in computer vision, research on crop disease recognition is presenting novel opportunities and challenges.

The masked autoencoder (MAE) model is one of the most representative MIM methods proposed by He et al.[32]. MAE does not require manually annotated datasets; it utilises the data itself for training and enables the model to learn high-level image features through mask reconstruction tasks. MAE has been adapted for crop disease detection research. For example, Zhang et al.[33] proposed an information entropy mask visual transformation model (IEM–ViT) for the rapid and accurate identification of tea diseases. This strategy integrates information entropy weighting with MAE, where the encoder processes only a subset of visible image patches, and the decoder reconstructs the masked patches by predicting the missing pixels, thereby facilitating parameter sharing and data enhancement. Experimental findings indicated that IEM–ViT achieved an accuracy of 93.78% in classifying seven tea diseases. Liu et al.[34] proposed a self-supervised transformer pre-training framework incorporating feature relationship conditional filtering (FRCF) and a latent semantic masking autoencoder (LSMAE) to optimise feature learning and representation.

Drawing on the aforementioned approaches, this study presents a novel detection model for tomato leaf disease. To enhance tomato leaf disease detection performance, the model employs MAE as the backbone, integrates an attention mechanism to learnable spatial relationships and employs a cross-layer feature fusion mechanism. The primary contributions of this study are summarised as follows:

(1) This research proposes a masked autoencoder pre-training model with learnable spatial relations, which leverages a learnable spatial weight matrix to enable each attention head to concentrate on specific spatial relationships. This design refines the model's ability to capture intricate structural information across different image regions, thereby boosting robustness in extracting fine-grained disease features.

(2) A cross-layer feature fusion mechanism is introduced, dynamically integrating and weighting features across encoder layers. The approach facilitates complementary information exchange across different feature levels, thereby enhancing the model's ability to capture the semantic details of subtle disease features and accurately distinguish among various disease types.

2 Materials and methods

2.1 Datasets

The tomato disease dataset of this study was derived from the AI Challenger 2018. Disease image samples were available for nine categories: yellow leaf curl virus, mosaic virus, target spot, spider mites, septoria leaf spot, leaf mould, late blight, early blight and bacterial spot. Healthy tomato image samples were also extracted. After applying data augmentation to diseases with limited original data, there are 14,259 tomato leaf images. Table 1 lists the quantities of each sample, and Fig. 1 displays representative images.

2.2 Methods

To address the challenges of identifying subtle disease spots and distinguishing between highly similar disease types in tomato leaf disease recognition, this study proposes a MAE pre-training method that incorporates spatial information via cross-layer feature fusion. The proposed method masked a majority of the image patches, using only the remaining visible patches to predict the missing portions, and applied an encoder–decoder architecture to complete the pre-training

Table 1 Quantity of each sample in tomato leaf disease dataset

Type of tomato diseases	Total quantity
Bacterial spot	1728
Early blight	1225
Late blight	1386
Leaf mold	1235
Septoria leaf spot	1496
Spider mites	1476
Target spot	1269
Yellow leaf curl virus	1780
Mosaic virus	1267
Healthy	1397
Total	14,259

task. This approach was designed to enhance the model's ability to capture local details in sophisticated backgrounds while optimising its generalisation performance across different disease types.

2.2.1 MAE

MAE is a form of denoising autoencoder (DAE)[35] with an asymmetric encoder–decoder design. The encoder generates latent representations of the image, whereas the decoder reconstructs the original image from these latent features. The loss is computed by comparing the reconstructed image output by the decoder with the original image. The original MAE uses a standard transformer as its encoder. The core idea is to mask a high proportion of an image's patches, encode only the unmasked patches and reconstruct the full image using the decoder. Figure 2 depicts the overall MAE framework.

The transformer encoder comprises multiple stacked transformer blocks, each consisting of a multi-head self-attention (MSA) layer and a multilayer perceptron (MLP) layer, as illustrated in Fig. 3. For each element in the input sequence, the standard self-attention mechanism computes a weighted sum of all values in the sequence. This mechanism, a core component of the transformer model, generates outputs by calculating relationships between input elements. The following equation governs the self-attention mechanism:

$$\text{Attention}(Q,K,V)=\text{SoftMax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where Q represents the query matrix, K denotes the key matrix and V is the value matrix. The term d_k refers to the dimensionality of the key vectors and is used to scale the dot product to prevent excessively large values, which could lead to the vanishing gradient problem. The first step involves computing the dot product of the query matrix Q and the key matrix K, resulting in the attention score matrix, which is shown in the following equation:

$$e_{ij}=Q_iK_i \tag{2}$$

To prevent numerical instability, the computed attention scores are typically scaled as follows:

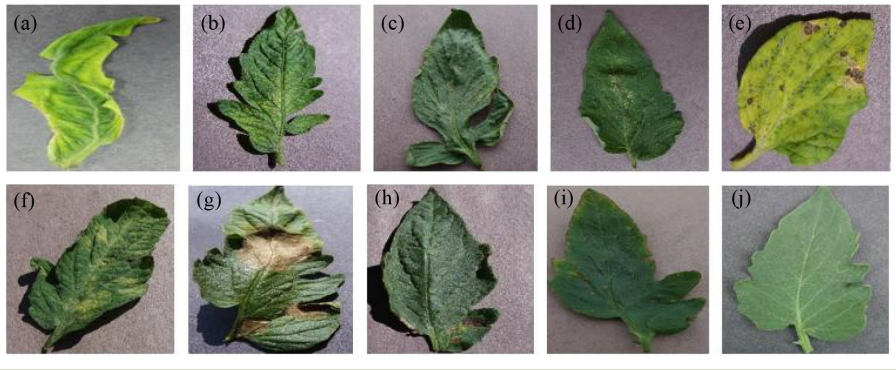


Fig. 1 Sample image of tomato leaf disease dataset: (a) yellow leaf curl virus spot; (b) mosaic virus; (c) target spot; (d) spider mites; (e) septoria leaf spot; (f) leaf mold; (g) late blight; (h) early blight; (i) bacterial spot; (j) healthy. The photos are sourced from the AI Challenger 2018 dataset, and are made available under an open-source Creative Commons license.

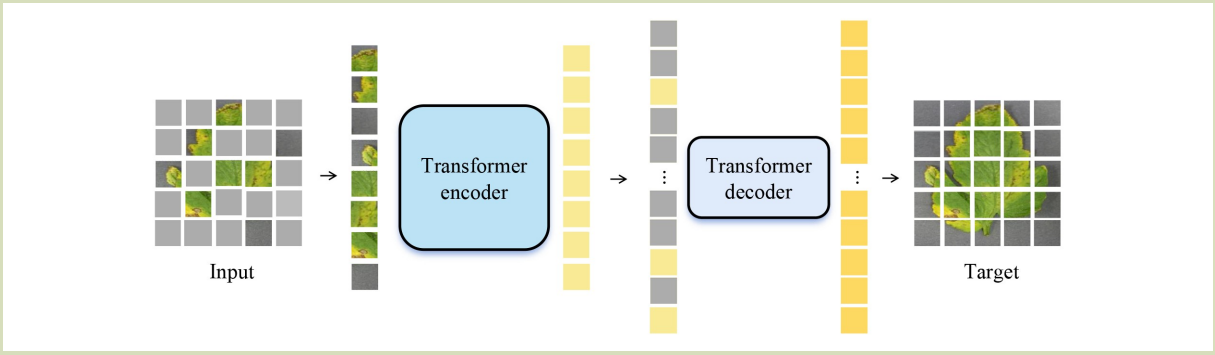


Fig. 2 Framework of MAE.

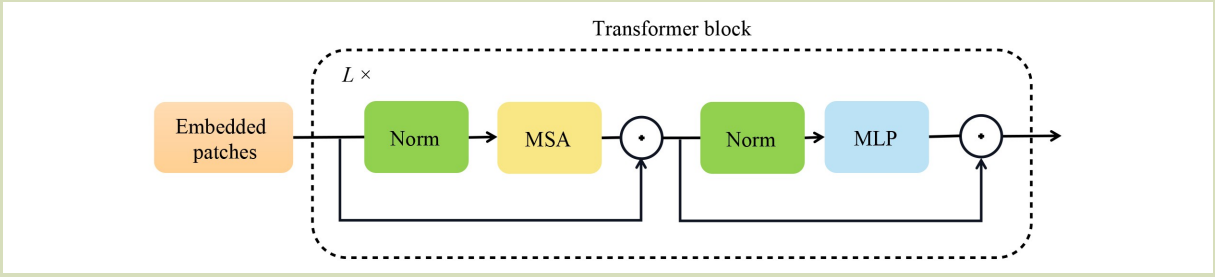


Fig. 3 Framework of transformer block.

$$\text{Scaled score}_{ij} = \frac{e_{ij}}{\sqrt{d_k}} \tag{3}$$

The scaled attention scores are then passed through a SoftMax function to produce a probability distribution over positions, representing the importance of each position. Finally, a weighted sum is computed using these probabilities, which is presented in the following equations:

$$\text{Attention score}_{ij} = \frac{\exp(\text{Scaled score}_{ij})}{\sum_{k=1}^{n} \exp(\text{Scaled score}_{ik})} \tag{4}$$

$$\text{Output}_i = \sum_{j=1}^{n} \text{Attention score}_{ij} V_j \tag{5}$$

2.2.2 CSF–MAE model network structure

The overall framework of CSF–MAE is depicted in Fig. 4, which combines an attention mechanism for learnable spatial relationship module (LSRM) and a cross-layer feature fusion mechanism. The model training process is divided into two parts: pre-training and fine-tuning.

(1) Pre-training process

The unmasked portion of the image was input into the encoder (which comprises multiple LSRM encoder layers, as shown in Fig. 4), and a specific set of two-dimensional spatial relationships was highlighted in each attention head to enable the model to calculate attention scores in the context of such spatial relationships, enhancing the model's understanding of spatial structure. After extracting features from the visible regions, the encoder performed cross-layer feature fusion by selecting shallow, intermediate and output layers. By combining features from multiple levels, the model can better adapt to convoluted, variable agricultural environments and optimise disease recognition accuracy. The fused feature representations were combined with the masked portions of the image and passed to the decoder. The decoder reconstructed the masked regions of the image from the feature representations generated by the encoder. During decoding, the model reconstructed the full image, which was then compared with the original image to compute the loss. The loss function is shown as follows:

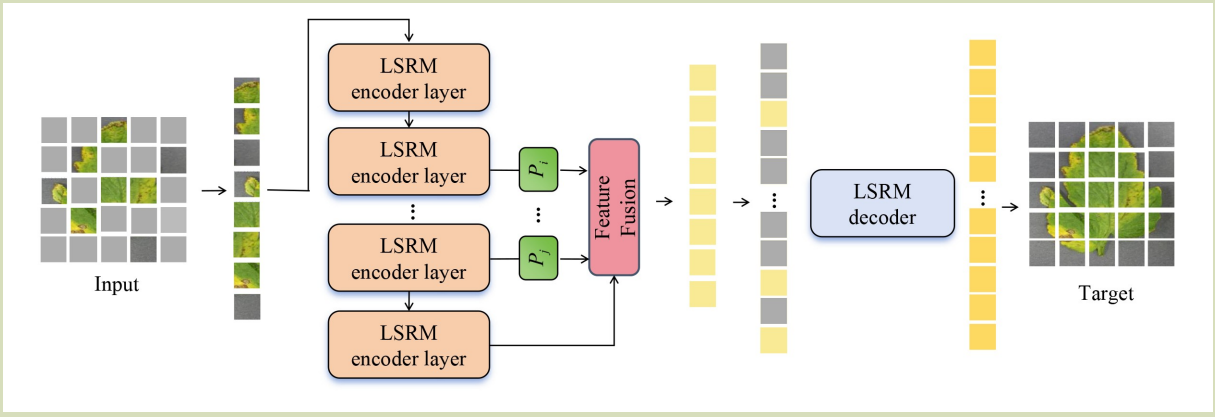


Fig. 4 Framework of CSF-MAE.

$$\mathcal{L}_{\mathrm{MAE}} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2 \tag{6}$$

where x_i represents the true value of the ith pixel in the original image, and $\hat{x}_i$ is the predicted value of the corresponding pixel in the model reconstruction output.

(2) Fine-tuning process

Fine-tuning is a commonly used technique in deep learning, where a pre-trained model is further trained to adapt to novel tasks or datasets. During self-supervised pre-training, the model learns image representations through proxy tasks. The fine-tuning process builds on this foundation by removing the decoder from the pre-trained architecture and adding a novel classifier on top of the encoder for downstream tasks.

Two typical approaches to fine-tuning are full fine-tuning and linear probing. Compared to fine-tuning, linear probing involves fewer trainable parameters. Because the backbone network constitutes the majority of the model and does not participate in backpropagation, linear probing offers faster training and shorter runtime but typically results in lower accuracy. In crop disease classification tasks, models often must learn highly specific features that may differ from those learned during pre-training. Full fine-tuning updates all model parameters across layers, enabling the model to better adapt to the specific task and dataset. Therefore, employing full fine-tuning was considered to optimise the model for crop disease classification. Figure 5 illustrates that the pre-trained encoder was retained and a specially designed classification head replaced the decoder during pre-training. The classification head was connected to the encoder's back end for classification,

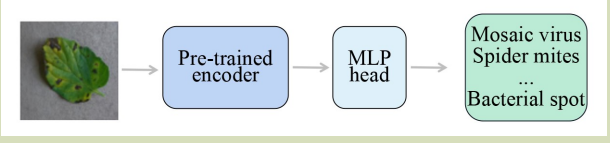


Fig. 5 Fine-tuning stage flowchart.

and the entire network's weights were updated via fine-tuning to effectively extract image features and improve model performance in crop disease classification.

2.2.3 Learnable spatial relationship module

Although self-attention mechanisms are effective at capturing long-range dependencies in input data, they often overlook the inherent spatial structure of images, leading to suboptimal performance while processing local features. In tomato leaf disease recognition, small disease spots or early symptoms are often subtle and challenging to identify. The task is further complicated by common agricultural environmental factors such as lighting variations, shadow effects and other sources of interference. To address these challenges and better focus on local features, reduce background interference, and enhance the detection and classification of early-stage diseases, this study designed a LSRM architecture with learnable spatial relations. This design highlights specific two-dimensional spatial relationships within each attention head, enabling the model to not only learn spatial priors similar to those induced by convolutions (focussing on local neighbourhoods) but also

capture long-range spatial relationships. As a result, the model can effectively extract both local and non-local features from images, enhancing its understanding of spatial structures.

The LSRM-based encoder processes only the visible, unmasked image patches and comprises several LSRM blocks (Fig. 6). Each block consists of a learnable spatial relations self-attention (LSR–SA) module and an MLP module. Layer normalisation was applied before each block, and residual connections were used after each block to stabilise the training process and accelerate convergence.

The LSR–SA module extends the traditional self-attention mechanism. Its core idea is to incorporate spatial relationships into the calculation of attention scores. Taking the query patch i as a reference point, the relative coordinates of image patch j were determined. A shared mapping function f_p, referred to as the spatial relations function, was applied to all query-key patch pairs. f_p was parameterised using a multilayer perceptron to enable the model to learn different types of spatial relationships. The formulas are represented in the following equations:

$$\Omega_{ij} = f_p(\vec{r}_{ij}) \tag{7}$$

$$f_p(\vec{r}_{ij}) = MLP(\vec{r}_{ij}) \tag{8}$$

This function maps relative coordinates to abstract scores, which highlight specific spatial relationships. In this way, the model learned various types of spatial relationships, including both local and non-local, and incorporated them into the attention computation. The output was a spatial matrix Ω_{ij}, which was combined with content-relevance scores to produce enhanced attention scores. These improved scores were then used for information aggregation. Each image was divided into $N \times N$ patches, and each $\Omega \in R^{N \times N}$ served as a spatial context for calculating the attention scores$A \in R^{N \times N}$, derived from the coordinate-based spatial relationships between input elements, specifically the relative positions of q and k. Under this attention mechanism, which ensured identical shapes for all attention scores, integration was performed via element-wise multiplication. The resulting scores were normalised using a SoftMax function, and the final output was obtained by a weighted sum of the value matrix V. The formulas are shown in the following equations:

$$\tilde{e}_{ij} = e_{ij} \cdot \Omega_{ij} \tag{9}$$

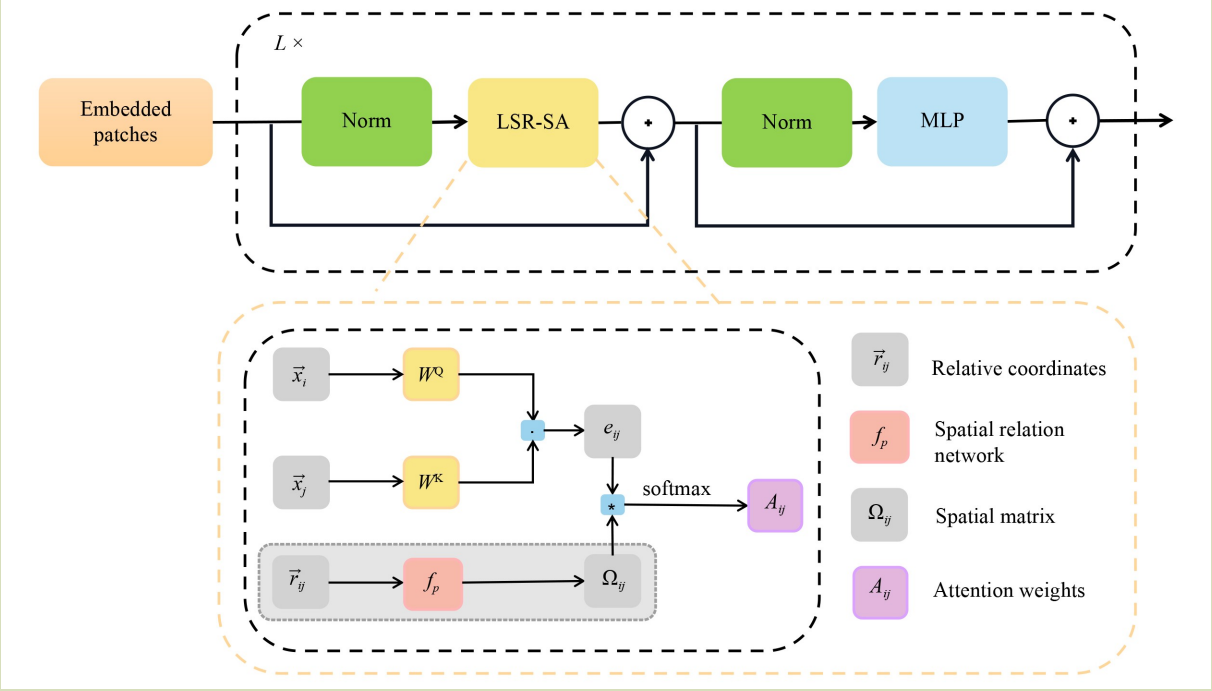


Fig. 6 Framework of LSRM block.

$$A_{ij} = \frac{\exp(\tilde{e}_{ij})}{\sum_{k=1}^{N} \exp(\tilde{e}_{ik})} \tag{10}$$

$$y_i = \sum_{j=1}^{N} A_{ij}(x_i W_v) \tag{11}$$

A multi-head attention mechanism was adopted to further enhance the model's ability to capture different types of dependencies. Each head independently computed attention scores, and the outputs of all heads were concatenated. The concatenated outputs were then transformed through a linear layer. The formula is shown as follows:

$$\text{Multi} - \text{Head}\ LSR - SA(Q,K,V) = Concat(y_1, \ldots, y_i, \ldots, y_n)W^O \tag{12}$$

where y_i represents the output of the ith head, and WO represents the linear transformation matrix.

By introducing LSRM to capture spatial relationships at different scales, the model can not only identify subtle disease spots on leaves but also perceive the health status of entire plants and broader regions. This capability is critical for assessing the overall health condition of crops.

After processing the visible image block output in the overall encoder comprising multiple LSRM encoder layers, the feature representation token obtained was combined with the remaining masked image blocks to create a complete set. This set was then passed to a lightweight LSRM decoder, which reconstructed the masked part of the image. In this way, the model can not only learn the relationships between different local regions in the image, but also enhance its understanding of the global structure, thereby effectively extracting disease features.

2.2.4 Cross-layer feature fusion

In computer vision tasks, conventional pixel-based autoencoder models can extract critical image features but often rely excessively on low-level features for pixel reconstruction, limiting their ability to capture high-level semantic information. Different levels of feature information have different semantics and meanings. By integrating different levels of feature information to assist pixel reconstruction, the characteristics of crop diseases can be more comprehensively captured. Tomato leaf disease images, the focus of this study, present challenges due to complex backgrounds affected by varying lighting conditions, shooting angles and crop growth stages. In addition, shallow features rich in local details may gradually diminish as successive transformations propagate into deeper layers. To address this issue, this study introduces a cross-layer fusion method to ensure that the model fully utilises information across layers, thereby enhancing generalisation across different disease types and improving recognition accuracy.

To avoid redundancy caused by multi-level fusion, this study strategically selected a subset of shallow, deep and partially intermediate layers for fusion. Shallow layers captured edges and textures, intermediate layers blend semantic and structural information, and deep layers encoded high-level abstract semantics. The framework of the cross-layer feature fusion module is shown in Fig. 7.

First, features were extracted from different layers of the encoder. Given an input image $I \in R^{H\times W\times 3}$, it was passed through the encoder E and obtained latent representations after multiple LSRM layers. The formula is shown in the following equation:

$$X = E(I) \tag{13}$$

For the feature extraction blocks within the encoder, the output features of the ith block were denoted as x_i, and the latent representations were $X = \{x_0, x_1, \ldots, x_{N-2}, x_{N-1}\}$, where N represents the encoder's depth, which is the number of LSRM layers. Selecting intermediate layers involves a large search space and can be challenging. To simplify this process and avoid optimisation challenges caused by redundant intermediate layers, this study ultimately selects four additional intermediate layers, excluding the shallow and output layers, resulting in a total of six layers. The output features of these selected layers were used for fusion.

The indices of these selected layers are defined as L. Before fusion, a projection layer P_i is applied to each selected layer to mitigate domain or distribution differences across layers. The formula is shown in the following equation:

$$\tilde{X} = \{P_i(x_i)\}_{i\in L} + \{x_{N-1}\} \tag{14}$$

Subsequently, a fusion layer F was introduced to integrate the multilayer features.

To ensure effective feature fusion across layers, a dynamic weight allocation strategy was employed. This strategy

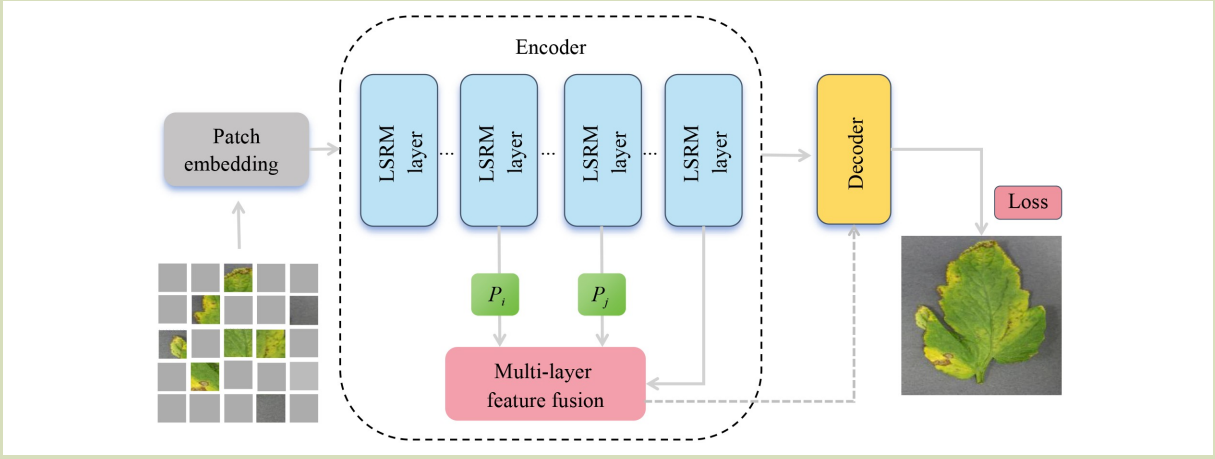


Fig. 7 The framework of the cross-layer feature fusion module.

automatically adjusted the weights of each layer based on the importance and relevance of its features. The specific steps are described as follows: each layer was initially assigned a weight $w_i^{(0)}$, which was dynamically updated to w_i during back propagation, allowing for a better integration of features across layers. Next, element-wise addition was performed, as shown in the following equation:

$$\text{Output} = \sum_{i \in L} w_i P_i(x_i) + w_{N-1} x_{N-1} \tag{15}$$

where w_i represents the weight of each selected fusion layer and w_{N-1} corresponds to the weight of the output layer. During training, the weights are dynamically updated using an optimiser while ensuring that their sum equals 1. This constraint ensured the stability and reliability of the fusion results. The fused features were then applied to reconstruct the original pixel values of the masked patches.

2.2.5 Model evaluation indicators

The main evaluation indicators for image classification include accuracy, precision (P), recall (R) and *F*1-score; the formulas are shown in Eqs. (16)–(19). The higher the value of each evaluation indicator, the better the network-trained model is.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

$$P = \frac{TP}{TP + FP} \tag{17}$$

$$R = \frac{TP}{TP + FN} \tag{18}$$

$$F1\text{-score} = 2 \times \frac{P \times R}{(P + R)} \tag{19}$$

where TP refers to the samples that are actually positive classes and correctly identified as positive classes by the model, FP reflects the samples that are actually negative but incorrectly identified as positive by the model, TN refers to the samples that are actually negative and correctly identified as negative by the model, and FN represents the samples that are actually positive but incorrectly identified as negative by the model.

2.3 Experimental environment and dataset production

All experiments in this study were conducted on an Ubuntu 20.04 operating system with a hardware configuration including an Nvidia GeForce RTX 3080 GPU and PyTorch as the deep learning framework.

The tomato leaf disease dataset was randomly split into a training set and a test set at 8:2. The training set was used for model training and parameter adjustment. In contrast, the testing set was used for the independent evaluation of the model's generalisation ability. During the specific training process, the data preprocessing involved re-sizing all images to 224 × 224 pixels, followed by data augmentation techniques such as random horizontal flipping and random cropping to enhance the model's generalisation ability. The specific experimental parameters are defined as follows: during pre-training, a 70% image masking ratio was used, with a batch size of 16. The optimiser employed was Adam W, with an initial

learning rate of 1e−3 and weight decay set to 0.05. A total of 400 epochs were trained, with the encoder having a depth of 12 and the decoder a depth of 8. During the subsequent fine-tuning phase, the pre-trained model was used for classification. The optimiser, learning rate and weight decay settings from the pre-training phase were retained, and the model was fine-tuned for 100 epochs.

3 Results

3.1 Determination of masking ratio

For MAE, higher masking ratios can considerably lower the amount of data the encoder needs to process, making the training of large models more efficient. As the masking ratio increases, the model must infer more content from less information, which increases the difficulty of the pre-training task. However, this also forces the model to learn more abstract and generalised feature representations. To test appropriate masking ratios, performance tests were conducted on the basic MAE model using the aforementioned datasets at different masking ratios, and the results are shown in Fig. 8.

Figure 8 reveals that the model's reconstruction performance initially improves with increasing masking ratio, then saturates. Among all the tested masking ratios, the tomato leaf disease datasets achieved the highest recognition accuracy at 70%. This finding suggests that a 70% masking ratio strikes an ideal balance. It ensures efficient training while encouraging the model to learn more robust and generalised feature representations, ultimately improving its overall performance in disease recognition tasks. Therefore, the masking ratio of this research model is 70%.

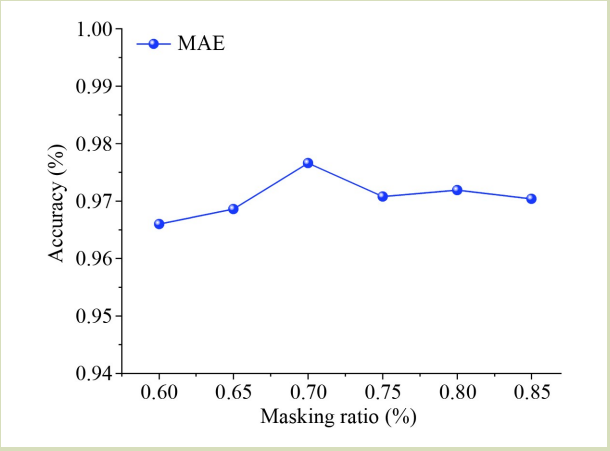


Fig. 8 Classification accuracy of MAE at different masking ratios.

3.2 Comparative analysis with different networks

3.2.1 Evaluation metrics of different models in the validation set

To validate the superiority of the CSF–MAE method in crop disease recognition tasks, it was compared with MAE and various improved MAE methods (including MixMAE[36], CrossMAE[37], MR–MCMAE[38] and CL–MAE[39]). The optimised MAE methods used official public implementations with recommended hyperparameters wherever available.

Table 2 shows that CSF–MAE not only outperformed MAE but also surpassed improved MAE methods in terms of classification accuracy, precision, recall and *F*1-score. Compared to MixMAE, CrossMAE, MR–MCMAE, CL–MAE, and MAE, accuracy improved by 0.37%, 0.73%, 0.18%, 0.66% and 1.39%, respectively. Precision improved by 0.32%, 0.71%, 0.13%, 0.69%, and 1.33%, respectively. Recall improved by

Table 2 Recognition results of different models on tomato leaf disease dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	*F*1-score (%)
MAE	97.66	97.68	97.66	97.66
MixMAE	98.68	98.69	98.66	98.67
CrossMAE	98.32	98.30	98.26	98.27
MR-MCMAE	98.87	98.88	98.87	98.87
CL-MAE	98.39	98.32	98.32	98.31
CSF-MAE	99.05	99.01	99.00	99.00

0.34%, 0.74%, 0.13%, 0.68% and 1.34%, respectively. *F*1-score showed improvements of 0.33%, 0.73%, 0.13%, 0.69% and 1.34%, respectively.

These results demonstrated CSF–MAE's ability to accurately extract local details and distinguish highly similar disease categories, even in complex backgrounds. CSF–MAE not only demonstrated consistent performance improvements across all datasets but also highlighted its robustness and adaptability to different crop disease types and environmental conditions. This outcome is attributed to the introduction of LSRM and cross-layer feature fusion modules in the CSF–MAE model. The LSRM module increases the model's sensitivity to subtle differences by promoting the learning of local features. The cross-layer feature fusion module effectively integrates multi-level information, ensuring the consistency and integrity of global features. This multi-level feature fusion strategy enables CSF–MAE to capture information at different levels, effectively learning essential abstract features and generalisation ability for crop diseases, thereby upgrading the model's overall performance.

In addition, the computational complexity and inference speed of different models on the tomato leaf disease dataset, as measured by indicators Params, FLOPs and FPS, are listed in Table 3.

Table 3 reveals that the number of parameters in this study's model has not increased significantly compared to the original MAE and the enhanced MAE. However, this method's accuracy, precision, recall and *F*1-score were higher than theirs (Table 2).

3.2.2 The discriminative ability of network models

To demonstrate the classification and recognition capabilities of the constructed CSF–MAE network model, the confusion matrices of five network models were compared (Fig. 9).

Figure 9 demonstrates that the CSF–MAE model has considerably greater discriminative ability than MAE, except for bacterial spot disease. For the four models, MixMAE, CrossMAE, MR–MCMAE and CL–MAE, CSF–MAE also performs better, although the margin of superiority is not as significant as that compared to MAE. The reason is that these four models are improved versions, and their performance is inherently good.

3.2.3 Performance compared with MAE

Compared with MAE and improved MAE methods, the experimental results for CSF–MAE showed higher accuracy and robustness. To visually demonstrate the performance differences, the accuracy, precision, recall and *F*1-score curves for MAE and CSF–MAE on the validation sets of the tomato disease datasets are shown in Fig. 10, providing a visual comparison of their performance.

From Fig. 10, CSF–MAE outperforms MAE on the tomato leaf disease dataset. Overall, the performance shows that as fine-tuning continues, the CSF–MAE accuracy, precision, recall and *F*1-score on tomato disease datasets gradually increase and eventually stabilise at high levels. This not only suggests the model's gradual adaptation to the data but also reflects its efficiency and stability in disease recognition tasks.

3.2.4 Visual results

To better demonstrate the model's effectiveness in processing input images, particularly its performance under high-occlusion conditions, this section presents the visual results of the model under high occlusion (Fig. 11).

Table 3 Computational complexity and inference speed of different models on tomato leaf disease dataset

Method	Params (M)	FLOPs (G)	FPS
MAE	111.908	21.982	80.216
MixMAE	114.666	17.860	45.439
CrossMAE	111.908	9.428	89.764
MR-MCMAE	115.041	16.896	86.272
CL-MAE	147.501	16.406	79.243
CSF-MAE	118.405	10.196	101.634

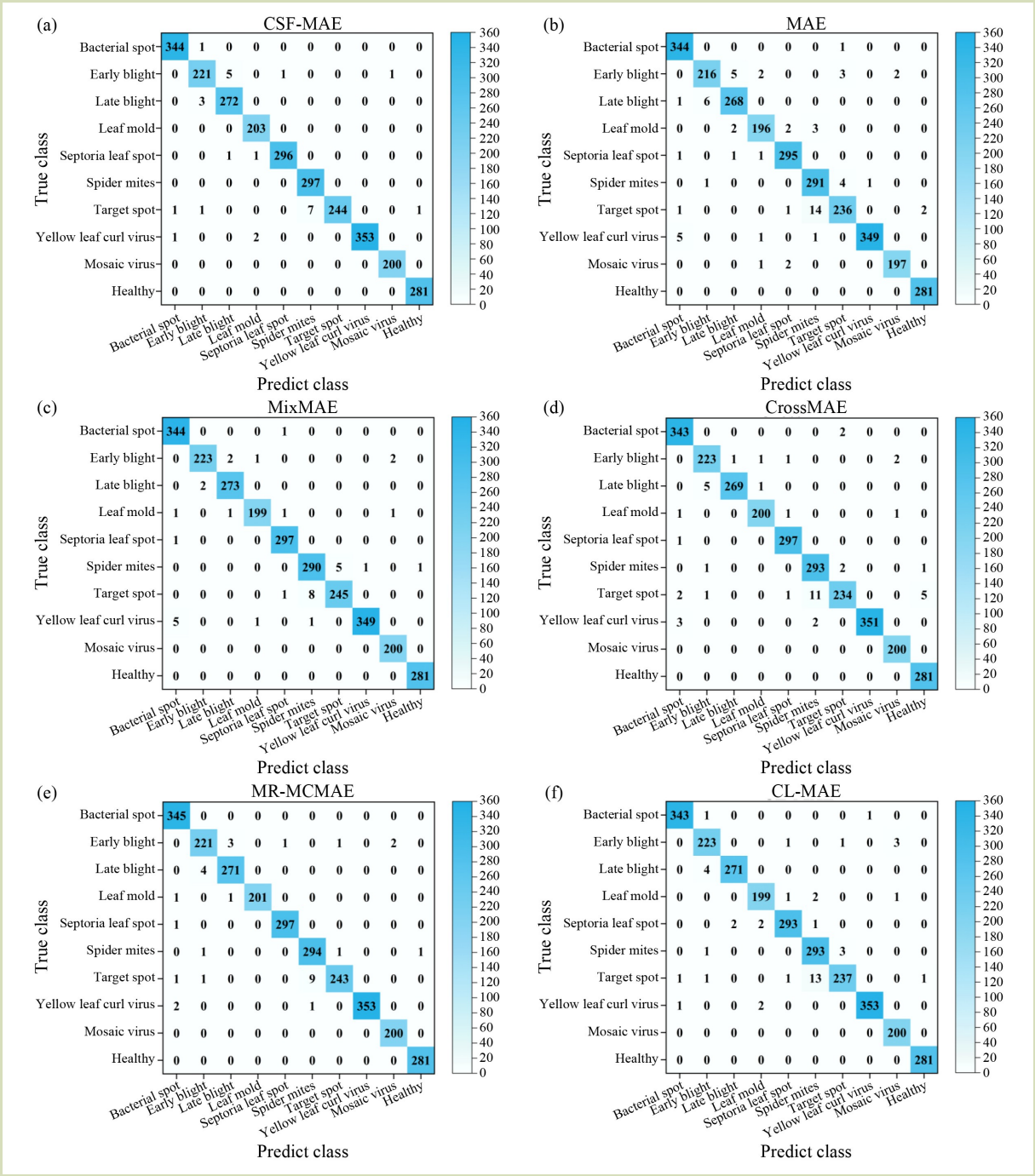


Fig. 9 The confusion matrices of six network models: (a) CSF-MAE; (b) MAE; (c) MixMAE; (d) CrossMAE; (e) MR-MCMAE; (f) CL-MAE.

Figure 11 illustrates the masking and reconstruction process of CSF–MAE, presenting three images for each example: the original, masked and reconstructed images. Unmasked patches are displayed in their original colours, whereas masked patches are displayed in grey to indicate occluded regions. Figure 11 shows that CSF–MAE can produce semantically coherent reconstructions even when 70% of the input image is randomly masked. This finding implies that the method not only assists

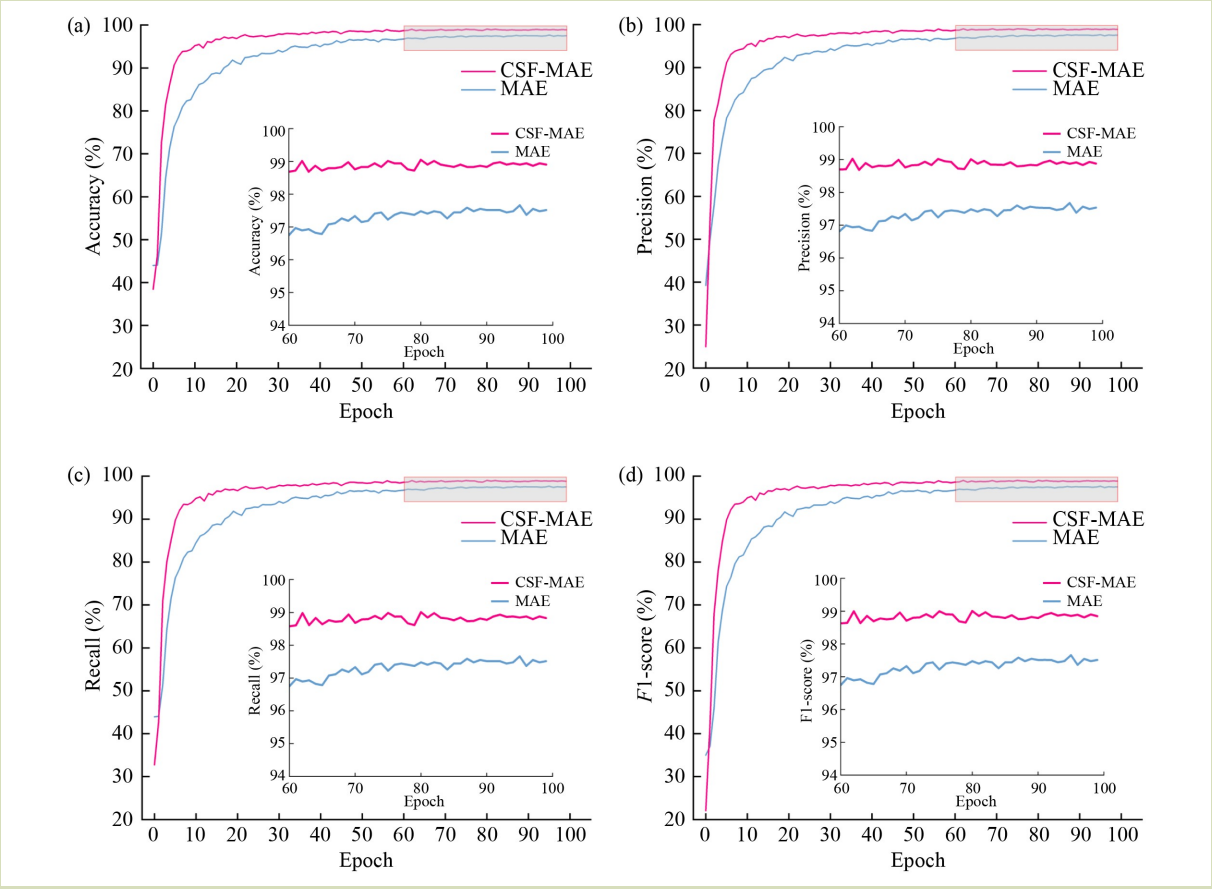


Fig. 10 Comparison of accuracy, precision, recall and *F*1-score between MAE and CSF-MAE: (a) accuracy; (b) precision; (c) recall; (d) *F*1-score.

the model in learning high-level feature representations but also effectively captures both local and non-local image features. In particular, the high masking ratio improves the model's ability to handle local details in intricate backgrounds and enhances its robustness in extracting subtle disease features. This capability is crucial for crop disease recognition, as diseases often manifest as fine-grained and complex textures or morphological changes, which are key to accurate diagnosis.

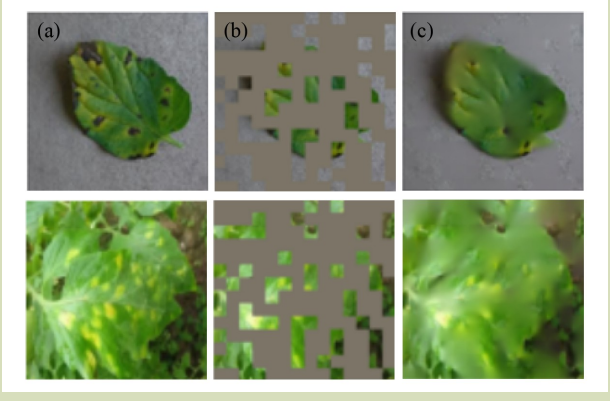


Fig. 11 Example results of CSF-MAE masking reconstruction image: (a) original image; (b) masked image; (c) reconstructed

4 Discussion

4.1 Advantages of CSF–MAE

The primary challenge of YOLO-based crop disease detection methods is the need for a large annotated dataset[40–43]. CSF–MAE not only does not depend on annotated data but also does not require all image information, using MIM, which enables the model to learn global image latent features and reduce image redundancy.

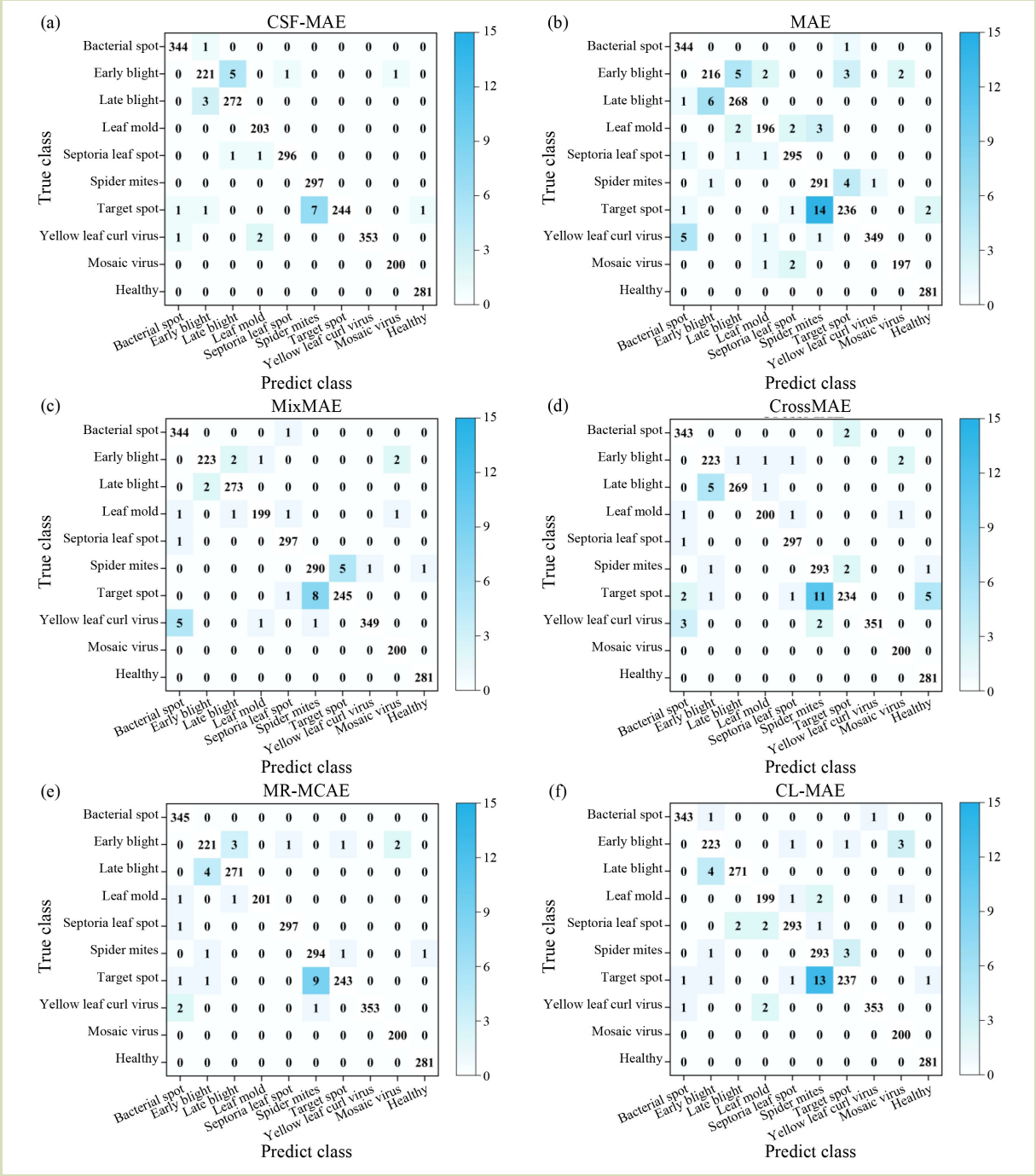


Fig. 12 The misclassification confusion matrices of six network models: (a) CSF-MAE; (b) MAE; (c) MixMAE; (d) CrossMAE; (e) MR-MCMAE; (f) CL-MAE.

From an algorithmic perspective, CSF–MAE's ability to learn both local and global features without relying on large-scale annotated samples stems from the synergy between the self-supervised reconstruction objective and the two structural designs, namely LSRM and cross-layer feature fusion. First, random high-ratio masking converts representation learning

into a pretext task that predicts masked patches from visible patches, allowing the model to learn lesion textures, edge patterns, and inter-region dependencies directly from unlabelled images rather than category annotations. Second, LSRM explicitly injects relative spatial relationships into self-attention, enabling different attention heads to focus on both local neighbourhoods around subtle lesions and long-range dependencies across distant regions, thereby simultaneously strengthening fine-grained local modelling and holistic structural perception. Third, the cross-layer feature fusion module dynamically integrates shallow texture and edge cues, intermediate structural information, and deep semantic representations, thereby jointly constraining masked-image reconstruction with local details and global semantics. Because the decoder must recover the masked regions from these fused encoder representations, the encoder is forced to learn latent features that are both locally discriminative and globally consistent, which explains why the proposed method remains effective even when annotated samples are limited.

4.2 Limitations and prospects

4.2.1 Limitations

Compared with the aforementioned model, CSF–MAE performs better than other models. However, CSF–MAE is only verified on public datasets, without considering datasets collected in the field. Applying methods to datasets collected in the field is a trend in computer technology applications in agriculture.

Furthermore, confusion matrix analysis revealed that misidentification primarily occurs between diseases due to inter-class similarity (Fig. 12). For instance, each model demonstrated significant difficulty in distinguishing between early blight and late blight (as shown in row 2 column 3 and row 3 column 2 of each confusion matrix in Fig. 12), two diseases that share similar visual symptoms, frequently leading to mutual misclassification. A comparable pattern of confusion was observed between spider mites and target spot (as shown in row 6 column 7 and row 7 column 6 of each confusion matrix in Fig. 12). Although CSF–MAE performed better than other models in identifying such cases, its errors were still concentrated in this category. Therefore, accurately distinguishing diseases with inter-class similarity is a critical issue for future research.

4.2.2 Prospects

One challenge in crop disease detection is the impact of inter-class similarity and intra-class variability[5]. Therefore, it is vital to reduce the effects of intra-class variability and inter-class similarity. Future work will focus on optimising the CSF–MAE model architecture, exploring additional feature fusion strategies, and expanding its applicability to accurately distinguish crop diseases with inter-class similarity or intra-class variability.

5 Conclusions

This study presented a novel self-supervised pre-training method based on MIM to enhance the precision and robustness of crop disease recognition. Unlike traditional contrastive SSL methods that rely heavily on data augmentation, CSF–MAE employs random masking to generate both visible and masked regions within images. This strategy avoids focussing solely on the main subject, ensuring all local regions can be occluded, and forces the encoder to capture high-quality feature representations from the visible portions, substantially enhancing its ability to learn comprehensive global features.

Acknowledgements

This work was funded by the Fujian Agriculture and Forestry University Technology Innovation Special Fund Project, China (117/KFb22098XA).

Compliance with ethics guidelines

Junyan Zhu, Enqi Yang, Chunji Shi, Changcai Yang, Liangquan Wu, and Yaohai Lin declare that they have no conflicts of interest or financial conflicts to disclose. This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Sharma D, Sharma A, Gautam H R. Green synthesized nanoparticles for disease management in vegetable crops: a review. *Plant Nano Biology*, 2025, **13**: 100179
2. Wang Q, Yan N, Qin Y S, Zhang X D, Li X. BED-YOLO: an enhanced YOLOv10n-based tomato leaf disease detection algorithm. *Sensors*, 2025, **25**(9): 2882
3. Wang Y, Zhang P X, Tian S. Tomato leaf disease detection based on attention mechanism and multi-scale feature fusion. *Frontiers in Plant Science*, 2024, **15**: 1382802
4. Yan C M, Li H L. CAPNet: tomato leaf disease detection network based on adaptive feature fusion and convolutional enhancement. *Multimedia Systems*, 2025, **31**(3): 178
5. Xiang Y Y, Gao S, Li X P, Li S Q. DWTFormer: a frequency-spatial features fusion model for tomato leaf disease identification. *Plant Methods*, 2025, **21**(1): 33
6. Nigam S, Jain R, Marwaha S, Arora A, Haque M A, Dheeraj A, Singh V K. Deep transfer learning model for disease identification in wheat crop. *Ecological Informatics*, 2023, **75**: 102068
7. Yao X T, Yang F, Yao J Y. YOLO-wheat: a wheat disease detection algorithm improved by YOLOv8s. *IEEE Access*, 2024, **12**: 133877–133888
8. Tegegne A G, Walle Y M, Haile M B, Yehulu G T, Yohannes S T. Comparative evaluation of CNN architectures for wheat rust diseases classification. *Discover Applied Sciences*, 2025, **7**(10): 1070
9. Haque A, Marwaha S, Deb C K, Nigam S, Arora A, Hooda K S, Soujanya P L, Aggarwal S K, Lall B, Kumar M, Islam S, Panwar M, Kumar P, Agrawal R C. Deep learning-based approach for identification of diseases of maize crop. *Scientific Reports*, 2022, **12**(1): 6334
10. Osouli S, Bolourian Haghighi B, Sadrossadat E. Mob-INC: an effective scheme for maize disease recognition based on deep networks. *Expert Systems with Applications*, 2026, **296**: 129006
11. Quach L D, Thai-Nghe N, Nguyen C N. Explainable dual-branch hybrid model for maize leaf disease classification. *SN Computer Science*, 2025, **6**(6): 585
12. Mahdi E T, AL-Kubaisy W J, Mahmood M. Capsule networks for rice leaf disease classification. *Journal of Intelligent Systems and Internet of Things*, 2025, **14**(2): 1–7
13. Wang B Y, Zhou H B, Xie H, Chen R L. Identification of rice disease based on MFAC-YOLOv8. *Journal of Real-Time Image Processing*, 2025, **22**(2): 75
14. Liu T L, Liu M G, Yang C C, Wu A C, Li X D, Wei W Z. Lightweight model improvement and application for rice disease classification. *Electronics*, 2025, **14**(16): 3331
15. Palei S, Mohapatra P. Metaheuristic-integrated DensoMobileForestNet for apple leaf disease identification. *Applied Fruit Science*, 2025, **67**(3): 142
16. Ahmad Wani O, Zahoor U, Ahmad Shah S Z, Khan R. Apple leaf disease detection using transfer learning. *Annals of Data Science*, 2025, **12**(1): 213–222
17. Dai D D, Liu H. An improved multiscale fusion dense network with efficient multiscale attention mechanism for apple leaf disease identification. *Frontiers of Agricultural Science and Engineering*, 2025, **12**(2): 173–189
18. Sun H, Fu R, Wang X W, Wu Y T, Al-Absi M A, Cheng Z Q, Chen Q, Sun Y M. Efficient deep learning-based tomato leaf disease detection through global and local feature fusion. *BMC Plant Biology*, 2025, **25**(1): 311
19. He Z Y, Tong M J. LT-YOLO: a lightweight network for detecting tomato leaf diseases. *Computers, Materials & Continua*, 2025, **82**(3): 4301–4317
20. Liu Z F, Guo X Y, Zhao T, Liang S. YOLO-BSMamba: a YOLOv8s-based model for tomato leaf disease detection in complex backgrounds. *Agronomy*, 2025, **15**(4): 870
21. Wang X W, Liu J. TomatoGuard-YOLO: a novel efficient tomato disease detection method. *Frontiers in Plant Science*, 2025, **15**: 1499278
22. Yang Z X, Xiao Y X, Zhang G M, Wang H M. MWD-YOLO11n: A Tomato Leaf Disease Detection Method. In: 2025 6th International Conference on Computer Engineering and Application. Hangzhou, China: *IEEE*, 2025, 292–296
23. Ye R, Zhang J, Gao Y J, Xu J Q, Gao Q. YOLOv8_LASKTR: A Multi-Dimensional Tomato Disease Identification Algorithm in Complex Environments. In: 2024 International Conference on Machine Intelligence and Digital Applications. Ningbo, China: *ACM*, 2024, 86–95
24. Lee Y S, Patil M P, Kim J G, Choi S S, Seo Y B, Kim G D. Improved tomato leaf disease recognition based on the YOLOv5m with various soft attention module combinations. *Agriculture*, 2024, **14**(9): 1472
25. Liu X Y, Lei H T, Zhou Y, Feng J M, Niu G, Zhou Y X. Tomato Leaf Disease Detection Based on Improved YOLOv8. In: 2024 6th International Conference on Internet of Things, Automation and Artificial Intelligence (IoTAAI). Guangzhou, China: *IEEE*, 2024, 145–150
26. Yang G F, Chen G P, He Y, Yan Z Y, Guo Y, Ding J. Self-supervised collaborative multi-network for fine-grained visual categorization of tomato diseases. *IEEE Access*, 2020, **8**: 211912–211923
27. Jin H B, Chu X Q, Qi J F, Zhang X X, Mu W S. CWAN: self-supervised learning for deep grape disease image composition. *Engineering Applications of Artificial Intelligence*, 2023, **123**: 106458
28. Kar S, Nagasubramanian K, Elango D, Carroll M E, Abel C A,

Nair A, Mueller D S, O'Neal M E, Singh A K, Sarkar S, Ganapathysubramanian B, Singh A. Self-supervised learning improves classification of agriculturally important insect pests in plants. *The Plant Phenome Journal*, 2023, **6**(1): e20079

29. He K M, Fan H Q, Wu Y X, Xie S N, Girshick R. Momentum Contrast for Unsupervised Visual Representation Learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: *IEEE*, 2020, 9726–9735
30. Dike H U, Zhou Y M, Deveerasetty K K, Wu Q T. Unsupervised Learning Based on Artificial Neural Network: A Review. In: *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)*. Shenzhen, China: *IEEE*, 2018, 322–327
31. Talaei Khoei T, Ould Slimane H, Kaabouch N. Deep learning: systematic review, models, challenges, and research directions. *Neural Computing and Applications*, 2023, **35**(31): 23103–23124
32. He K M, Chen X L, Xie S N, Li Y H, Dollár P, Girshick R. Masked Autoencoders Are Scalable Vision Learners. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: *IEEE*, 2022, 15979–15988
33. Zhang J H, Guo H L, Guo J, Zhang J. An information entropy masked vision transformer (IEM-ViT) model for recognition of tea diseases. *Agronomy*, 2023, **13**(4): 1156
34. Liu H L, Zhan Y Z, Xia H F, Mao Q R, Tan Y X. Self-supervised transformer-based pre-training method using latent semantic masking auto-encoder for pest and disease classification. *Computers and Electronics in Agriculture*, 2022, **203**: 107448
35. Vincent P, Larochelle H, Bengio Y, Manzagol P A. Extracting and Composing Robust Features with Denoising Autoencoders. In: *Proceedings of the 25th International Conference on Machine Learning*. Helsinki, Finland: *ACM*, 2008, 1096–1103
36. Liu J H, Huang X, Zheng J L, Liu Y, Li H S. MixMAE: Mixed and Masked Autoencoder for Efficient Pretraining of Hierarchical Vision Transformers. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: *IEEE*, 2023, 6252–6261
37. Fu L T, Lian L, Wang R H, Shi B F, Wang X D, Yala A, Darrell T, Efros A A, Goldberg K. Rethinking patch dependence for masked autoencoders. *Transactions on Machine Learning Research*, 2025
38. Gao P, Lin Z Y, Zhang R R, Fang R Y, Li H Y, Li H S, Qiao Y. Mimic before reconstruct: enhancing masked autoencoders with feature mimicking. *International Journal of Computer Vision*, 2024, **132**(5): 1546–1556
39. Madan N, Ristea N C, Nasrollahi K, Moeslund T B, Tudor Ionescu R. CL-MAE: Curriculum-Learned Masked Autoencoders. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, HI, USA: *IEEE*, 2024, 2480–2490
40. Xie H B, Ding J H. MSAM-YOLO: an improved YOLO v8 based on attention mechanism for grape leaf disease identification method. *Journal of Computational Methods in Sciences and Engineering*, 2025, **25**(3): 2874–2882
41. Zhou Z H, Yu L M. Lightweight grape leaf disease identification for complex field scenes——YOLO-grape model and its boundary-lesion coalesce optimization policies. *Signal, Image and Video Processing*, 2025, **19**(16): 1378
42. Chen Y, Li H B, Chen Y P. MGDS-YOLO: a multi-scale apple disease detection algorithm in complex environments. *Engineering Letters*, 2025, **33**(12): 4925–4936
43. Fang K, Zhou R, Deng N, Li C, Zhu X H. RLDD-YOLOv11n: research on rice leaf disease detection based on YOLOv11. *Agronomy*, 2025, **15**(6): 1266