

Article

FAF-Net: A Lightweight Frequency-Auxiliary Fusion Network for Plant Disease Classification in Natural-Scene Images

Chu Qing Zhao ¹ and Fang Ling Sun ^{2,*}¹ College of Horticulture, China Agricultural University, Beijing 100193, China; chuqingzhao@cau.edu.cn² College of Science, China Agricultural University, Beijing 100083, China

* Correspondence: flsun@cau.edu.cn

Abstract

Plant disease classification in natural scenes remains challenging because disease symptoms are often localized and imaging conditions are complex, including cluttered backgrounds, illumination variations, scale changes, and fine-grained inter-class similarities. To address these challenges, this study proposes FAF-Net, a frequency-aware fusion network with auxiliary supervision for plant disease classification in natural scenes. The proposed framework is built on EfficientNet-B3 and integrates three complementary strategies: CutMix augmentation, an FFT-based frequency branch, and a healthy/diseased auxiliary supervision branch. The RGB branch extracts spatial semantic features from natural-scene images, whereas the frequency branch converts the input image into a log-normalized Fourier magnitude spectrum and learns complementary texture representations. The auxiliary branch provides coarse-grained health-status supervision during training, encouraging the shared representation to capture disease-relevant features. Experiments were conducted on the PlantDoc dataset, which contains 2598 images from 27 healthy and diseased categories. Compared with the EfficientNet-B3 baseline, FAF-Net improved the classification accuracy from 69.49% to 74.58%, corresponding to a gain of 5.09 percentage points. Ablation results further indicate that CutMix, frequency-domain features, and auxiliary supervision provide complementary improvements. These results suggest that frequency-aware feature fusion and coarse-grained auxiliary supervision can enhance plant disease classification under natural-scene conditions.

Keywords: plant disease classification; computer vision; PlantDoc; FFT; auxiliary supervision

Academic Editor: Atsushi Mase

Received: 26 May 2026

Revised: 22 June 2026

Accepted: 23 June 2026

Published: 25 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Plant diseases remain a major constraint on crop productivity, product quality, and food security. According to the Food and Agriculture Organization of the United Nations (FAO), up to 40% of global crop production is lost annually due to plant pests and diseases, highlighting the importance of timely disease monitoring and effective plant health management [1]. Leaves are among the most informative organs for visual diagnosis because many foliar diseases first appear as visible changes in leaf color, lesion morphology, spot distribution, and texture patterns [2]. Conventional diagnosis mainly relies on expert visual inspection, which is labor-intensive, subjective, and difficult to scale for large production areas or real-time field monitoring [3]. Therefore, image-based leaf disease

recognition using low-cost mobile devices has become an important direction for intelligent crop health monitoring.

Early image-based approaches for plant disease recognition generally relied on digital image processing and traditional machine learning techniques. These methods usually extracted handcrafted color, shape, and texture descriptors from leaf images, and then used classifiers such as Support Vector Machines (SVMs) [4], genetic-algorithm-based segmentation and soft-computing approaches [5], and discriminant analysis [6] for disease classification. Although handcrafted features are interpretable, their performance strongly depends on feature design, image quality, and acquisition conditions. With the rapid development of artificial intelligence and computer vision, deep learning has become a mainstream approach for plant disease recognition [7].

Convolutional neural networks (CNNs) can automatically learn hierarchical feature representations from raw images, reducing the dependence on handcrafted feature engineering [8,9]. CNN-based methods have been widely used in image classification, object detection, semantic segmentation, and plant disease diagnosis [10]. Substantial progress has been achieved in CNN-based plant disease recognition. For example, Mohanty et al. applied AlexNet and GoogLeNet to the PlantVillage dataset and achieved high classification accuracy across multiple crop-disease categories [11]. Sladojevic et al. developed a CNN-based model for plant disease recognition using leaf images collected from the Internet [12]. Ferentinos evaluated several deep learning architectures on a large-scale plant disease dataset and reported strong classification performance [13]. Too et al. compared several fine-tuned deep models, including VGG, Inception, ResNet, and DenseNet, and demonstrated the importance of backbone selection for plant disease identification [14]. These studies confirm the effectiveness of deep learning for plant disease recognition, but many reported high-accuracy results were obtained from relatively clean or controlled datasets, where leaves are usually centered, backgrounds are simple, and disease symptoms are visually clear. Such image conditions differ substantially from practical natural-scene environments.

Natural-scene plant disease classification remains more difficult than classification under controlled imaging conditions. Field or web-collected images often contain cluttered backgrounds, non-uniform illumination, shadows, occlusions, multiple leaves, varying shooting angles, and large-scale variations [15]. These factors may cause a model to rely on background context, leaf morphology, or imaging artifacts rather than pathological symptoms. PlantVillage is one of the most widely used datasets for plant disease recognition, but most of its images were collected under controlled or semi-controlled conditions [16]. To reduce the gap between controlled datasets and practical agricultural scenarios, Singh et al. released PlantDoc, a natural-scene plant disease dataset containing 2598 web-collected images from 27 healthy and diseased plant categories [17]. Compared with controlled datasets, PlantDoc contains more complex backgrounds, larger intra-class variation, stronger inter-class similarity, and more severe sample imbalance. Therefore, classification on PlantDoc requires representations that can capture disease-relevant cues under complex and weakly controlled imaging conditions.

For such a setting, the backbone network should provide sufficient representation capacity while avoiding unnecessary computational complexity. The EfficientNet family balances network depth, width, and input resolution through compound scaling, achieving a favorable trade-off between accuracy and computational cost [18]. EfficientNet-B3 provides strong ImageNet performance with moderate model complexity, making it a suitable backbone.

However, backbone selection alone does not fully address the difficulty of learning subtle disease cues from natural-scene images. A standard RGB-based CNN mainly learns spatial-domain representations, which are effective for capturing global semantics, leaf

shape, color distribution, and visible lesion appearance. These features may be less sufficient when disease symptoms are small, blurred, partially occluded, or mixed with complex background textures. Frequency-domain information provides a complementary perspective for representing local texture variations and fine-grained visual details. Neural networks have been reported to exhibit a spectral bias toward low-frequency patterns during training, which means that lower-frequency components are often learned more easily than high-frequency components [19]. However, disease-related cues such as lesion boundaries, spot edges, powdery regions, and texture disruptions may be partly reflected in high-frequency image components [20]. If these fine-grained cues are not sufficiently captured, models may fail to distinguish visually similar disease categories.

Recent agricultural vision studies have explored this issue through frequency-domain or frequency-spatial representations. DWTFormer introduced a frequency-spatial fusion framework for tomato leaf disease identification, showing that wavelet-based frequency information can complement spatial features [21]. FA-Unet further combined Fourier transform with attention mechanisms to enhance disease-related frequency-domain features in fruit leaf disease images under complex backgrounds [22]. WaveLiteNet integrated a discrete wavelet transform with MobileNetV3 for lightweight tea leaf disease recognition [23]. WGA-YOLO further used wavelet-based frequency recalibration to preserve high-frequency edge and texture details for crop disease detection under field-like conditions [24]. These studies support the usefulness of frequency-domain information for representing lesion boundaries and local pathological textures. Nevertheless, many existing frequency-aware models are designed around relatively complex attention structures, or are developed for specific crop-disease scenarios. For public natural-scene classification benchmarks such as PlantDoc, image quality, background complexity, sample imbalance, and limited annotations jointly affect recognition performance. Therefore, this study introduces an FFT-based frequency branch as a lightweight complementary feature extractor rather than an additional large backbone. This branch extracts frequency-domain features from the Fourier magnitude spectrum and fuses them with EfficientNet-B3 spatial features, thereby incorporating lesion-related texture cues with limited additional architectural overhead.

Another difficulty is that image-level category labels provide limited guidance for learning disease-relevant regions. Many plant diseases share similar visual symptoms, and these symptoms may occupy only a small region of the image. Under such conditions, the model may benefit from an additional learning signal that emphasizes whether the leaf shows disease-related changes while distinguishing among specific disease categories. Auxiliary supervision introduces additional learning signals during training and can encourage the network to learn more discriminative representations beyond the primary classification objective [25]. Previous studies have used multi-label learning, plant-type information, disease severity labels, or multi-task prediction to improve disease-relevant representation learning [26,27]. For example, multi-label frameworks can jointly model plant type, disease category, and severity, while multi-task systems can learn species-related and disease-related features simultaneously. However, these strategies often require additional annotations, such as plant type, disease severity, species labels, or lesion-level information. In contrast, a healthy/diseased auxiliary task can be generated directly from the original class labels and can provide coarse disease-status guidance without extra manual annotation.

To clarify the position of FAF-Net relative to recent related studies, Table 1 summarizes representative works from four perspectives: natural-scene recognition, frequency-domain feature learning, lightweight modeling, and auxiliary or multi-task supervision.

Table 1. Positioning of FAF-Net relative to representative recent plant disease recognition studies.

Study	Main Idea
Salman et al. [28]	Addressed plant disease classification under in-the-wild conditions by combining a Vision Transformer backbone with a Mixture-of-Experts strategy.
Xiang et al. [21]	Proposed DWTFormer to fuse wavelet-based frequency features with spatial representations for tomato leaf disease identification.
Li et al. [22]	Proposed FA-Unet, which combines Fourier-domain feature enhancement with attention mechanisms for fruit leaf disease identification under complex backgrounds.
Yang et al. [23]	Proposed WaveLiteNet by integrating discrete wavelet transform with a lightweight CNN for tea leaf disease recognition.
Yang et al. [26]	Developed a multi-label framework to jointly identify plant type, leaf disease category, and disease severity.
Keceli et al. [27]	Designed a multi-task prediction system for plant disease and species recognition from leaf images. Targets weak disease-cue learning in natural-scene image-level classification by combining data-level regularization, frequency-domain texture complementation, and coarse-grained disease-status supervision.
Ours, FAF-Net	

As summarized in Table 1, many recent studies tend to emphasize one of these directions, whereas natural-scene image-level classification involves a coupled representation-learning problem.

Based on the above considerations, this study proposes FAF-Net, a frequency-aware auxiliary fusion network for natural-scene plant disease classification. FAF-Net organizes three strategies around the same representation-learning bottleneck. CutMix regularizes the model at the data level and reduces dependence on complete leaf appearance or background context. The FFT-based branch complements RGB spatial features with frequency-domain texture cues related to lesion boundaries and local visual disruptions. The healthy/diseased auxiliary task provides coarse-grained disease-status guidance using labels derived from the original categories, without requiring lesion-level, severity-level, or species-level annotations. The expected practical implication is to provide a lightweight image-level recognition framework that may support low-cost preliminary screening of plant health status in field or mobile-device scenarios, without requiring lesion-level annotations. The main contributions of this study are as follows:

1. We propose a frequency-aware auxiliary fusion framework for natural-scene leaf disease classification. Rather than treating frequency learning, data augmentation, and auxiliary supervision as isolated components, FAF-Net integrates them into a unified framework to improve the robustness of disease-related representation learning under complex imaging conditions.
2. We design a lightweight FFT-based frequency branch to complement RGB spatial features extracted by EfficientNet-B3. This branch converts the input image into a log-normalized Fourier magnitude spectrum and learns compact frequency-domain representations, which are expected to capture lesion boundaries, spot textures, and local visual disruptions that may be weakened in purely spatial CNN features.
3. We introduce a healthy/diseased auxiliary supervision task to guide the shared representation toward disease-status-aware features during training. This coarse-grained supervision provides an additional constraint for fine-grained category classification, especially when different disease classes show similar symptoms or when lesion regions are small in natural-scene images.
4. We evaluate FAF-Net on the public natural-scene PlantDoc dataset and analyze its effectiveness through ablation experiments, comparative evaluation, confusion-matrix analysis, and mixed PlantDoc/PlantVillage analysis. The results provide evidence that frequency-domain complementary features and coarse-grained auxiliary

supervision can improve plant disease classification under challenging natural-scene conditions.

2. Methodology

2.1. Datasets

This study used the publicly available PlantDoc dataset [17] as the primary benchmark for natural-scene plant disease classification. PlantDoc contains 2598 images covering 27 healthy and diseased plant categories, including 17 diseased classes and 10 healthy classes. Unlike controlled leaf-image datasets, PlantDoc images were collected from natural or web-based scenes and contain complex backgrounds, non-uniform illumination, scale variations, multiple leaves, and diverse shooting angles. These characteristics make PlantDoc suitable for evaluating the robustness of plant disease classification models under natural-scene conditions. Its public availability also facilitates reproducible evaluation and comparison under consistent data sources and experimental protocols.

The class distribution of PlantDoc is shown in Figure 1. The categories are clearly imbalanced, which may bias the model toward classes with larger sample sizes. Representative examples from PlantDoc are shown in Figure 2, showing the visual complexity and intra-class variability of natural-scene plant disease images.

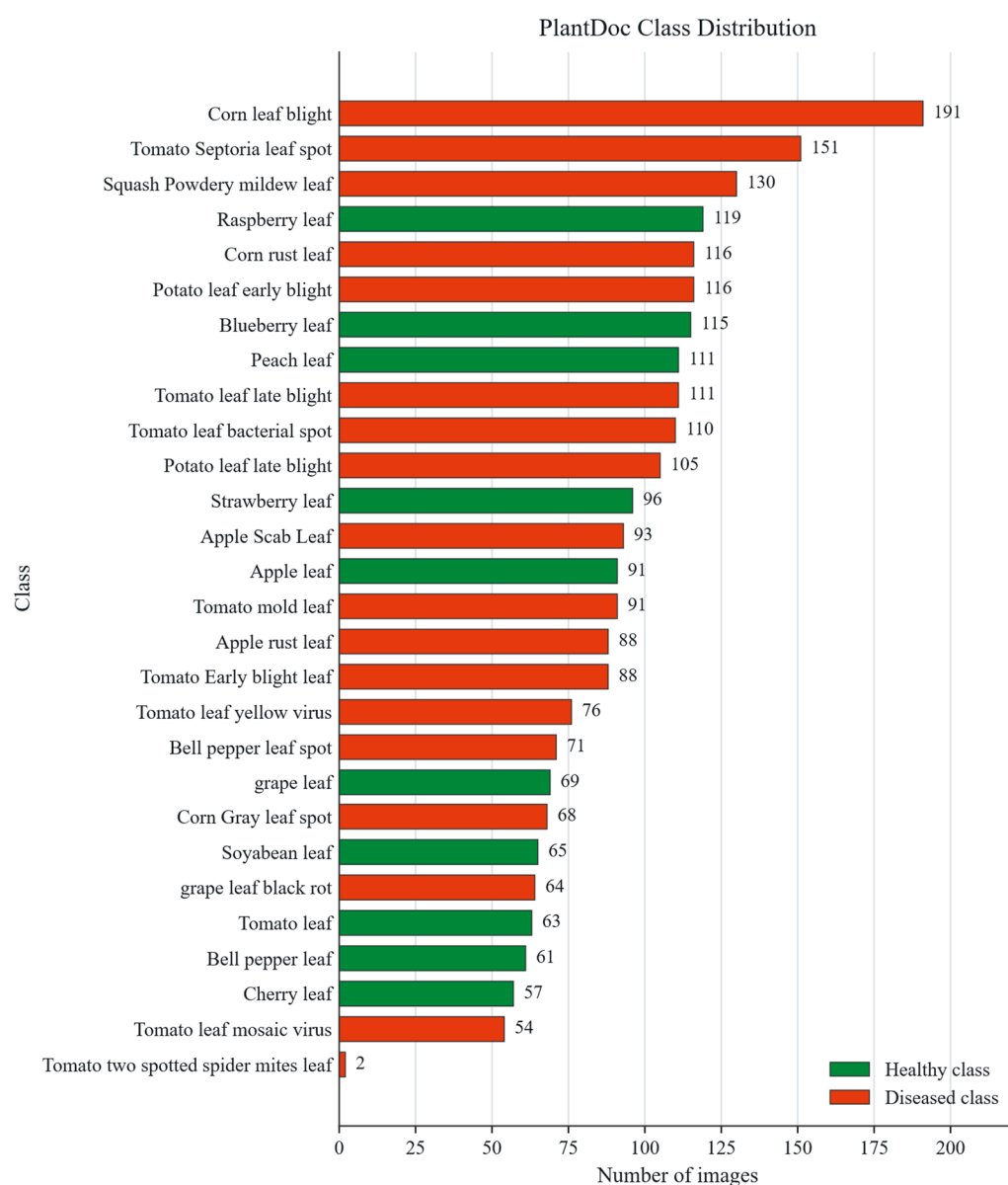


Figure 1. Class distribution of PlantDoc dataset. The dataset contains 2598 natural-scene images from 27 healthy and diseased plant leaf categories. Categories are sorted by the number of images to highlight the class imbalance among different crop-disease classes.

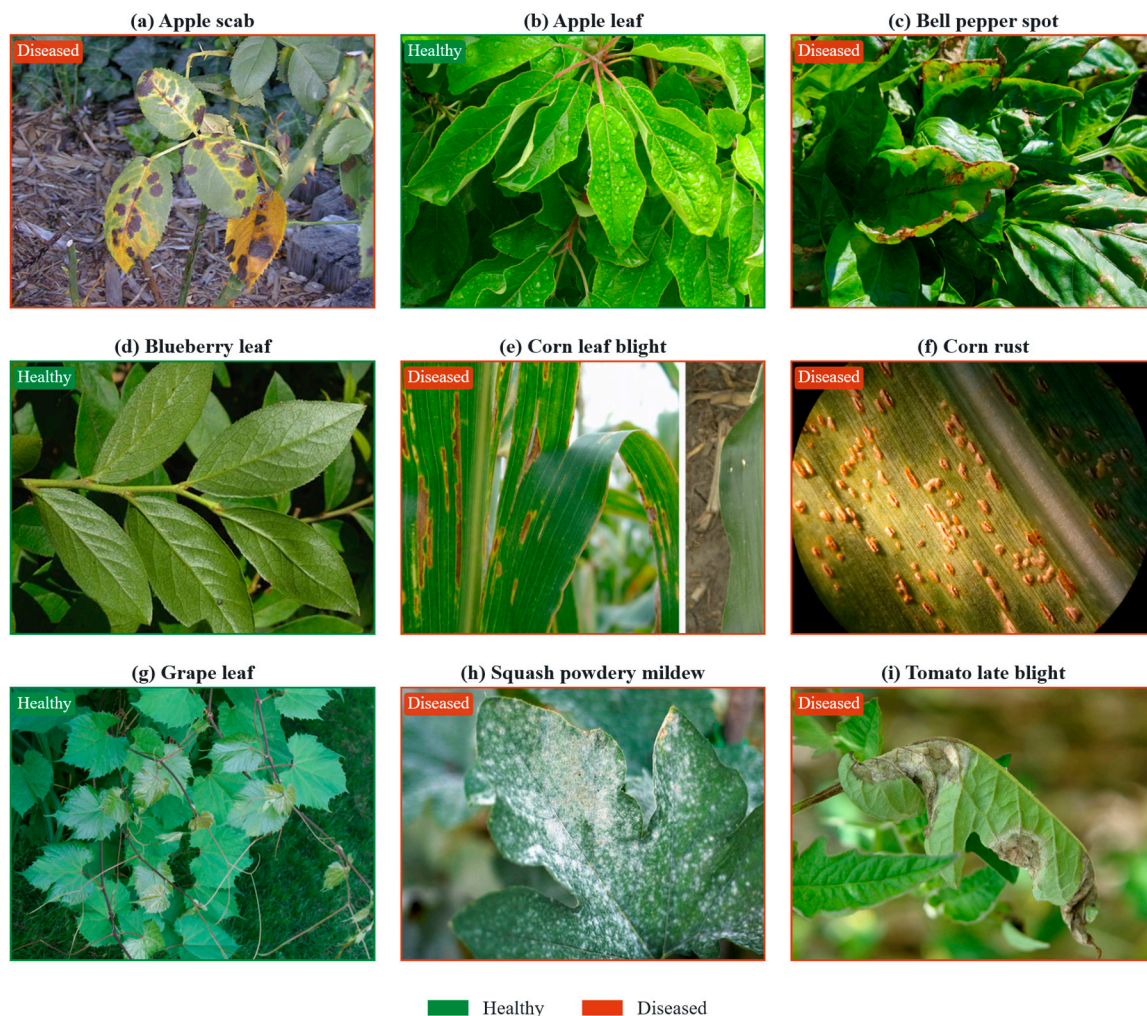


Figure 2. Representative images from PlantDoc dataset. The examples illustrate natural-scene plant disease images with complex backgrounds, non-uniform illumination, different lesion scales, and substantial visual variability.

2.2. Overview of FAF-Net

Figure 3 illustrates the overall architecture of the proposed FAF-Net. The model consists of four main components: an RGB spatial branch, an FFT-based frequency branch, a feature fusion module, and two classification heads. The RGB branch uses EfficientNet-B3 as the backbone to extract spatial semantic features from natural-scene images. In parallel, the frequency branch converts the input image into a Fourier magnitude spectrum and extracts complementary texture-related representations. The spatial and frequency-domain features are then fused and used for multi-class plant disease classification.

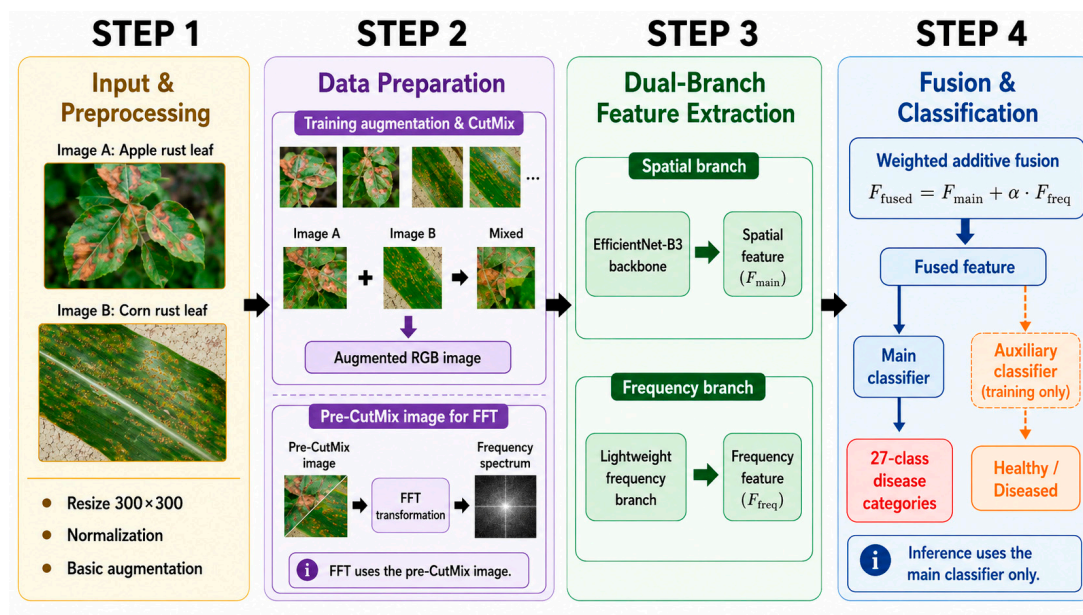


Figure 3. Model overview of FAF-Net. The model integrates an EfficientNet-B3 RGB branch and an FFT-based frequency branch. Spatial and frequency-domain features are fused for multi-class plant disease classification, while a healthy/diseased auxiliary classifier provides additional supervision during training.

During training, CutMix [29] augmentation is applied to improve the robustness of the model to localized occlusion, lesion variability, and background interference.

Let $\mathbf{x}$ denote the input image, and basic preprocessing procedures are first applied to obtain a preprocessed image $\mathbf{x}_p$. The RGB spatial branch receives the CutMix-augmented image $\tilde{\mathbf{x}}$ during training and extracts spatial semantic features using EfficientNet-B3. During inference, CutMix is not applied, and the RGB branch receives the preprocessed image $\mathbf{x}_p$. The RGB branch extracts a spatial feature vector $\mathbf{f}_s$ using EfficientNet-B3:

$$\mathbf{f}_s = E_{\theta}(\mathbf{x}_{\text{RGB}}), \quad (1)$$

where $E_{\theta}(\cdot)$ denotes the EfficientNet-B3 backbone. Here, $\mathbf{x}_{\text{RGB}}$ is set to $\tilde{\mathbf{x}}$ during training, whereas $\mathbf{x}_{\text{RGB}}$ corresponds to $\mathbf{x}_p$ during inference. In parallel, the FFT-based frequency branch extracts a compact frequency feature vector $\mathbf{q}_f$:

$$\mathbf{q}_f = G_{\phi}(\mathbf{x}_p), \quad (2)$$

where $G_{\phi}(\cdot)$ denotes the frequency feature extraction branch. The frequency feature is projected to the same dimensionality as the spatial feature:

$$\mathbf{f}_f = P_{\psi}(\mathbf{q}_f), \quad (3)$$

where $P_{\psi}(\cdot)$ denotes the linear projection layer. The final fused representation is obtained by learnable weighted fusion:

$$\mathbf{f} = \mathbf{f}_s + \alpha_f \mathbf{f}_f, \quad (4)$$

where α_f is a learnable coefficient initialized to 0.1. The fused feature $\mathbf{f}$ is then fed into two classification heads. The main classifier predicts the 27 PlantDoc categories, including both healthy and diseased plant classes. In addition, a healthy/diseased auxiliary classifier is introduced during training to provide coarse-grained disease-status supervision. This auxiliary task regularizes the shared representation and encourages the model to focus on disease-relevant visual cues. During inference, the final plant disease category is

determined only by the main classifier, while the auxiliary classifier is not used for final decision-making.

2.3. Data Preprocessing

All images were resized to 300×300 pixels. Several data augmentation methods were introduced in the training set. The main operations used in the preprocessing stage are as follows:

1. **Color Jittering:** Randomly adjusts brightness, contrast, saturation and hue. The adjustment ranges for brightness, contrast, and saturation were set to 0.2, and the hue adjustment range was set to 0.1.
2. **Random Horizontal Flipping:** Generates mirrored images by horizontally flipping the input images, introducing variations in leaf orientation and spatial position.
3. **Random Cropping and Resizing:** Randomly crops local regions of the image and resizes them back to 300×300 pixels, enabling the model to learn disease features at different scales. The cropping scale was set to 0.75–1.0, meaning that the cropped region occupied 75% to 100% of the original image area.

It should be noted that no leaf segmentation or background removal was performed during preprocessing. Instead, the complete natural-scene images were preserved as the model input.

2.4. CutMix-Based Augmentation

In addition to the conventional augmentation operations described in Section 2.3, CutMix was further adopted during training. CutMix is a data augmentation strategy proposed by Yun et al. [29], in which a rectangular region of one training image is replaced with the corresponding region cropped from another image. This strategy increases sample diversity and encourages the model to learn discriminative features from partial disease symptoms rather than relying only on a single global appearance pattern.

Given two training samples $(\mathbf{x}_{p_i}, \mathbf{y}_i)$ and $(\mathbf{x}_{p_j}, \mathbf{y}_j)$, where $\mathbf{x}_{p_i}$ and $\mathbf{x}_{p_j}$ are input images and $\mathbf{y}_i$ and $\mathbf{y}_j$ are their one-hot class labels, CutMix generates a mixed image $\tilde{\mathbf{x}}$ and a mixed label $\tilde{\mathbf{y}}$ as follows:

$$\tilde{\mathbf{x}} = \mathbf{M} \odot \mathbf{x}_{p_i} + (1 - \mathbf{M}) \odot \mathbf{x}_{p_j}, \quad (5)$$

$$\tilde{\mathbf{y}} = \lambda_{\text{adj}} \mathbf{y}_i + (1 - \lambda_{\text{adj}}) \mathbf{y}_j, \quad (6)$$

where $\mathbf{M} \in \{0,1\}^{H \times W}$ is a binary mask indicating the retained and replaced image regions, $\odot$ denotes element-wise multiplication, and λ_{adj} is the adjusted mixing ratio calculated according to the actual pasted area. The initial mixing ratio λ is sampled from a Beta distribution:

$$\lambda \sim \text{Beta}(\alpha_{\text{cm}}, \alpha_{\text{cm}}), \quad (7)$$

where α_{cm} is the CutMix hyperparameter, which is set to 1. If the pasted region has width r_w and height r_h , the adjusted mixing ratio is calculated as:

$$\lambda_{\text{adj}} = 1 - \frac{r_w r_h}{WH} \quad (8)$$

The mixed image $\tilde{\mathbf{x}}$ is then used as the input to the RGB spatial branch during training, whereas the FFT-based frequency branch receives the corresponding pre-CutMix preprocessed image to avoid artificial spectral artifacts introduced by CutMix boundaries. During inference, CutMix is not applied, and the preprocessed image is used directly as the input. An example of the CutMix operation used in this study is shown in Figure 4.

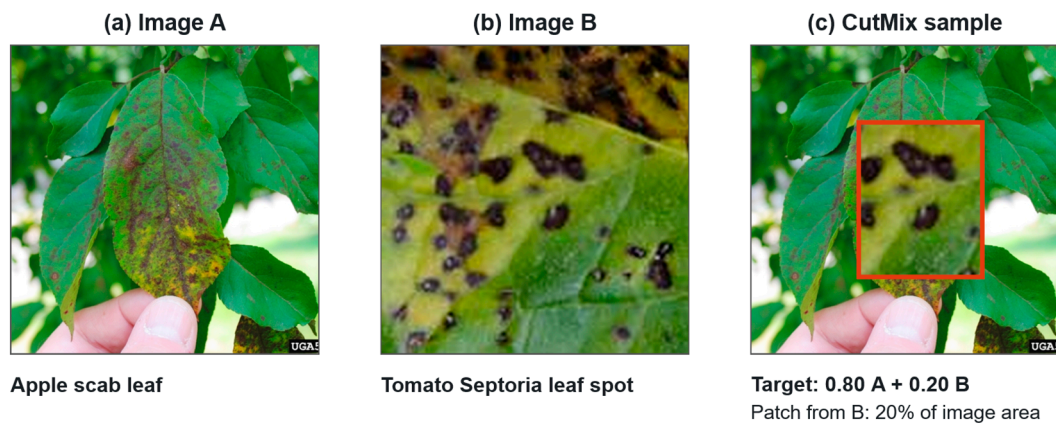


Figure 4. CutMix mixed sample.

2.5. FFT-Based Frequency Branch

The FFT-based frequency branch is designed to extract complementary frequency-domain information from the input image. Disease symptoms such as lesion boundaries, spot edges, powdery regions, and local texture disruptions may introduce frequency-domain variations that are not fully captured by spatial RGB features alone. Therefore, an additional frequency branch is introduced to enhance the representation of fine-grained texture cues.

To construct the frequency branch, the preprocessed RGB image $\mathbf{x} \in \mathbb{R}^{3 \times H \times W}$ is first converted into a grayscale image $\mathbf{x}_g$. A two-dimensional Fast Fourier Transform (FFT) is then applied to $\mathbf{x}_g \in \mathbb{R}^{H \times W}$ to obtain the complex frequency representation

$$\mathbf{X}_f = \mathcal{F}(\mathbf{x}_g), \quad (9)$$

where $\mathcal{F}(\cdot)$ denotes the 2D FFT operation. Since the FFT output is complex-valued, its magnitude spectrum is utilized to capture the frequency characteristics. To align the zero-frequency component with the center of the spectrum, a frequency shift operation $\text{fftshift}(\cdot)$ is executed. The centered magnitude spectrum A_f is defined as follows:

$$A_f = |\text{fftshift}(\mathbf{X}_f)|. \quad (10)$$

To reduce the large dynamic range of the spectral amplitudes, we apply a logarithmic scaling transformation

$$A'_f = \log(1 + A_f). \quad (11)$$

Finally, A'_f is normalized to the range of [0, 1] using Min-Max normalization, and used as the input to the frequency branch.

The frequency branch is designed with a lightweight convolutional architecture, comprising three successive Convolution-Batch Normalization-ReLU (Conv-BN-ReLU) blocks followed by a Global Average Pooling (GAP) layer. This pipeline extracts a 64-dimensional frequency feature, which is subsequently projected by a linear layer to match the dimensionality of the spatial feature extracted by the EfficientNet-B3 backbone.

2.6. Auxiliary Supervision Branch

A healthy/diseased auxiliary supervision branch is introduced to provide coarse-grained disease-status guidance during training [25]. The main classification task predicts the 27 PlantDoc categories, whereas the auxiliary task predicts whether the input image belongs to a healthy or diseased class. This auxiliary task is expected to regularize the shared feature representation and encourage the model to focus on disease-relevant visual cues.

The binary auxiliary labels were generated from the original PlantDoc category labels. Categories corresponding to pathological symptoms, including scab, rust, spot, blight, mildew, rot, virus, mold, and mite-related symptoms, were assigned to the diseased class. Categories describing normal crop leaves were assigned to the healthy class.

The auxiliary classifier uses the fused feature $\mathbf{f}$ as its input. While the main classifier predicts plant leaf category, the auxiliary classifier is implemented as a fully connected layer that predicts healthy/diseased status. During training, the model is optimized using a weighted combination of two classification losses.

Let L_{cls} denote the main classification loss, and L_{aux} denote the auxiliary binary classification loss, the total loss is defined as:

$$L = L_{cls} + \lambda_{aux}L_{aux}, \quad (12)$$

where λ_{aux} is the weight of the auxiliary loss. In this study, λ_{aux} is set to 0.2. This encourages the model to learn both fine-grained and coarse-grained information. During inference, the final plant disease category is determined only by the main classifier.

3. Experimental Setup

3.1. Implementation Details

The experiments were conducted on a high-performance computing platform equipped with an NVIDIA A100 GPU (80 GB), an 8-core Intel Xeon CPU, and 32 GB RAM, using Ubuntu 20.04, PyTorch, CUDA 11.4.0, and Python 3.8.

All input images were resized to 300×300 pixels before being fed into the network. During training, the basic augmentation operations described in Section 2.3 were applied to all models, including color jittering, random horizontal flipping, and random resized cropping. CutMix augmentation was further applied only to the corresponding model variants, as described in Section 2.4.

The models were optimized using the Adam optimizer. The initial learning rate was set to 6×10^{-4} , the batch size was set to 96, and all models were trained for 100 epochs. The CutMix parameter α_{cm} was set to 1. For FAF-Net, the auxiliary loss weight λ_{aux} was set to 0.2. The learnable frequency fusion coefficient α_f was initialized to 0.1 and updated jointly with the network parameters during training.

For a fair comparison, all ablation models and comparative baseline models were trained using the same data split, input resolution, basic augmentation strategy, optimizer type, and evaluation metrics. The final classification results were obtained from the main 27-class classification branch. The auxiliary healthy/diseased branch was used to provide additional supervision during training and was not used to determine the final plant disease category during inference.

The checkpoint with the highest validation macro-F1 score was selected for final testing. If two checkpoints produced the same validation macro-F1 score, the checkpoint with the higher validation accuracy was selected. The final evaluation was conducted on the official PlantDoc test set.

3.2. Evaluation Metrics

Accuracy, precision, recall and F1-score were used to evaluate the performance of the model. During evaluation, the predicted class was determined by selecting the class with the highest logit from the main classification branch.

Accuracy measures the overall proportion of correctly classified samples in the test set, and is defined as:

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}}, \quad (13)$$

where $N_{correct}$ denotes the number of correctly predicted samples, and N_{total} denotes the total number of test samples.

Since the PlantDoc dataset contains imbalanced class distributions, we further calculated precision, recall and F1-score for each class. For the i -th class, these metrics are defined as:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}, \quad (14)$$

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i}, \quad (15)$$

$$F1_i = \frac{2 \times \text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}, \quad (16)$$

where TP_i , FP_i and FN_i represent the true positives, false positives, and false negatives of the i -th class, respectively.

The final precision, recall, and F1-score were reported using macro averaging:

$$\text{Precision}_{macro} = \frac{1}{C} \sum_{i=1}^C \text{Precision}_i \quad (17)$$

$$\text{Recall}_{macro} = \frac{1}{C} \sum_{i=1}^C \text{Recall}_i \quad (18)$$

$$F1_{macro} = \frac{1}{C} \sum_{i=1}^C F1_i \quad (19)$$

where C is the total number of classes. Therefore, macro-averaged metrics provide a balanced evaluation of model performance on imbalanced datasets.

3.3. Computational Efficiency Evaluation

To evaluate the computational efficiency of FAF-Net, an additional benchmark was conducted under a unified testing protocol. The benchmark was performed on an NVIDIA GeForce RTX 3060 Laptop GPU with a batch size of 1. The external RGB input size was fixed at $1 \times 3 \times 300 \times 300$ for all models. The evaluated metrics included the number of parameters, multiply-accumulate operations (MACs), estimated floating-point operations (FLOPs), model size, inference latency, throughput, and peak GPU memory consumption.

The number of parameters was counted directly from the model parameters, while MACs were measured using THOP under the same 300×300 input resolution. FLOPs were estimated as two times the number of MACs. For latency measurement, each model was switched to evaluation mode and tested without gradient calculation. After 30 warm-up iterations, the average latency was calculated from 200 repeated forward passes. Throughput was reported as frames per second based on the average latency.

For all compared methods, the reported model latency refers to the neural network forward pass. For frequency-branch models, the model-forward benchmark used an additional precomputed $1 \times 1 \times 300 \times 300$ frequency input. In addition, the complete FAF-Net inference pipeline was also tested by including GPU-based FFT magnitude preprocessing before the network forward pass. Image reading, decoding, and resizing were not included in the latency measurement.

4. Results

4.1. Overall Experimental Results

Figure 5 shows the validation accuracy and macro-F1 curves of the EfficientNet-B3 baseline and the proposed model variants during training. The curves were used to monitor the training process and select the best checkpoint, whereas the final performance was evaluated on the official PlantDoc test set. Compared with the baseline, the models incorporating the proposed strategies generally maintained higher accuracy and macro-F1 values during training, indicating that the additional modules improved the stability of feature learning under natural-scene conditions.

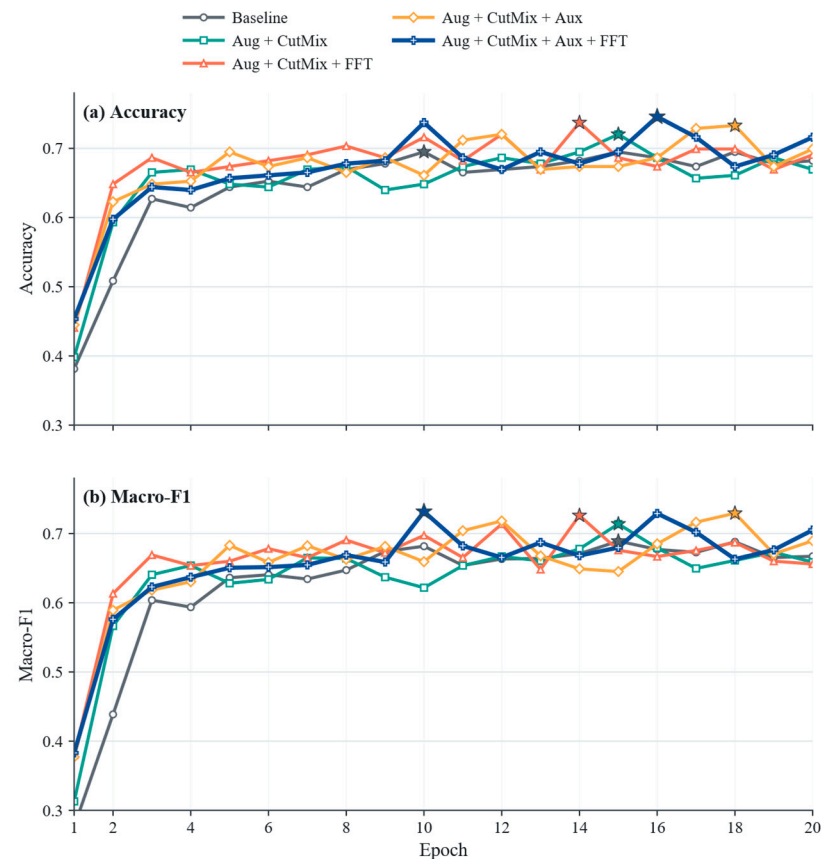


Figure 5. F1-score and accuracy. Star markers indicate the epoch at which each model achieved its highest accuracy in panel (a) or highest macro-F1 in panel (b). The curves were used for checkpoint selection.

The EfficientNet-B3 baseline achieved an accuracy of 69.49%. After introducing CutMix, FFT branch and auxiliary supervision branch, the overall performance improved substantially. The final model achieved 74.58% accuracy. Compared with the baseline model, the proposed methods improved accuracy by 5.09 percentage points. Our proposed methods can effectively improve the recognition ability of models on natural-scene images.

4.2. Ablation Study

To evaluate the contribution of each component, we gradually added CutMix, FFT branch and auxiliary classification supervision. The ablation results are shown in Table 2 and Figure 6.

Table 2. Ablation results of different FAF-Net variants on the PlantDoc test set. “✓” indicates that the corresponding component is included in the model variant, whereas “–” indicates that it is not used. Aux denotes the healthy/diseased auxiliary supervision branch.

Method	CutMix	FFT	Aux	Accuracy	Macro Precision	Macro Recall	Macro F1-Score
Baseline	–	–	–	0.6949	0.7107	0.6880	0.6818
Baseline + CutMix	✓	–	–	0.7203	0.7317	0.7255	0.7139
Baseline + CutMix + FFT	✓	✓	–	0.7373	0.7707	0.7335	0.7262
Baseline + CutMix + Aux	✓	–	✓	0.7331	0.7556	0.7282	0.7294
Baseline + CutMix + Aux + FFT	✓	✓	✓	0.7458	0.7465	0.7402	0.7291

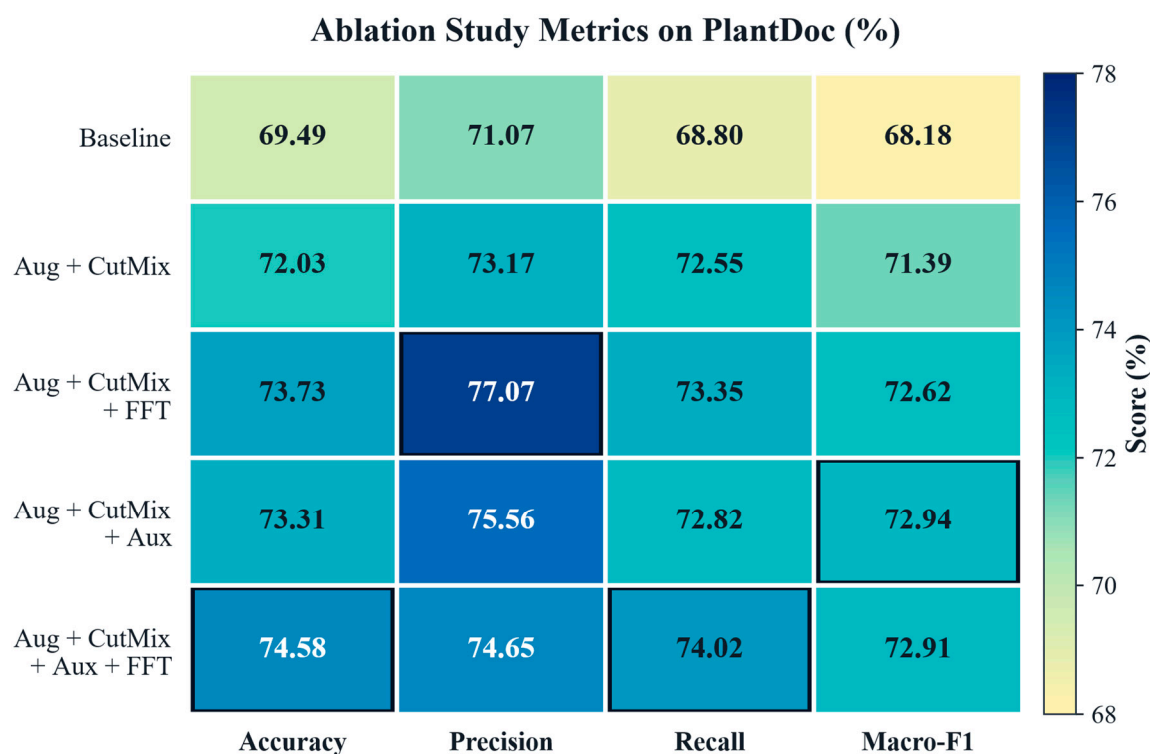


Figure 6. Visual comparison of ablation results across four evaluation metrics. The values in each cell are reported as percentages, and the color intensity indicates the relative performance level according to the color scale.

After adding CutMix as an extra data augmentation method, accuracy increased from 69.49% to 72.03%, and F1-score increased from 68.18% to 71.39%. CutMix increases sample diversity and improves model robustness to complex backgrounds and localized lesions.

When the FFT frequency branch was added on top of CutMix, accuracy further increased to 73.73%. After adding the FFT branch, the CutMix + FFT variant achieved the highest macro-precision, suggesting that frequency-domain features may have helped reduce false-positive predictions in some classes. In this context, a false positive means that a sample not belonging to a certain class is incorrectly predicted as that class. A higher macro-precision indicates that when the model predicts a certain class, the prediction is generally more reliable, meaning that the model makes fewer incorrect positive predictions for that class.

With auxiliary classification supervision added on top of CutMix, the model achieved the highest F1-score of 72.94%. Healthy/Diseased auxiliary supervision provides an additional constraint for shared feature learning and improves overall discriminative ability. This task encourages the model to first learn whether disease-related visual cues are

present and then further distinguish the specific 27 fine-grained categories. Therefore, after adding auxiliary supervision, the model became more capable of retrieving disease samples that might otherwise be missed. In other words, auxiliary supervision made the model more sensitive to potential disease-related features, rather than being overly conservative and predicting only the easiest samples.

Figure 7 shows the performance improvement of each module compared with Baseline + CutMix. The final model combined CutMix, the FFT branch, and auxiliary classification supervision, achieving the highest accuracy of 74.58% and recall of 74.02%. Three strategies are complementary in terms of data augmentation, frequency feature supplementation, and training supervision.

Performance Improvement on Ablation Study

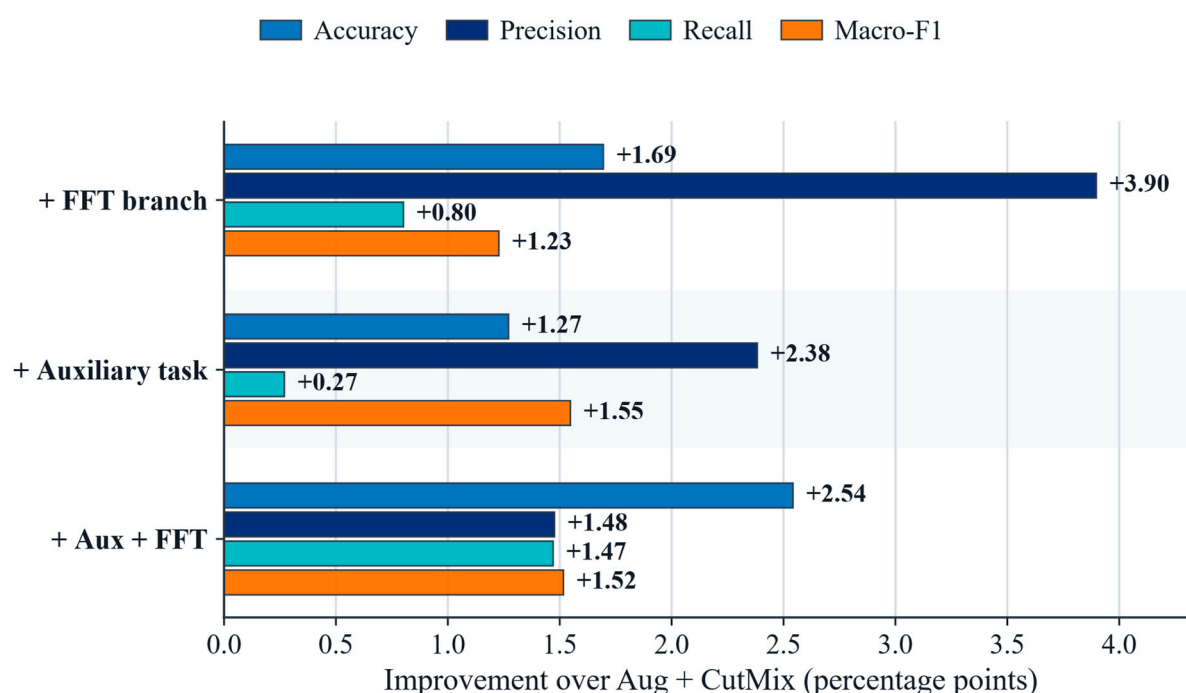


Figure 7. Performance improvement on ablation study.

As Figure 7 displays, the CutMix + FFT variant produced the highest macro-precision. Macro-precision reflects the reliability of positive predictions for each class. In other words, it measures how many samples predicted as a given class are actually from that class, and then averages this value across all classes. After adding healthy/diseased auxiliary supervision, the complete FAF-Net achieved the highest accuracy and macro-recall, but its macro-precision slightly decreased compared with its variant. This decrease can be explained by the change in prediction behavior introduced by auxiliary supervision. The auxiliary task encourages the shared representation to become more sensitive to general disease-related cues. As a result, the model can retrieve more difficult or ambiguous diseased samples, which improves recall and overall accuracy. However, this increased sensitivity may also cause the model to assign some visually similar samples to disease categories more actively.

For example, samples with similar spots, necrotic regions, or edge discoloration may be more easily predicted as a related disease class even when they belong to another class. These additional incorrect positive assignments increase false positives for certain categories, which can reduce macro-precision. Therefore, the slight decrease in macro-precision

does not indicate that the complete model is less effective; rather, it reflects a precision–recall trade-off, showing that FAF-Net improves class coverage and overall correctness.

4.3. Comparative Experiment and Analysis

To verify the effectiveness of the proposed method, VGG16, ResNet50, DenseNet121, and InceptionV3 were selected as comparison models. All models were evaluated under the same PlantDoc split and trained under the same basic settings. ImageNet-pretrained weights and basic data augmentation strategies were applied to all models.

As shown in Table 3, compared to other models, DenseNet121 achieved an accuracy of 69.07%, establishing a stable performance. The selected EfficientNet-B3 baseline performed slightly better than DenseNet121. FAF-Net's accuracy is higher by 5.09 percentage points compared with EfficientNet-B3. FAF-Net achieved the highest accuracy, Macro-Recall, and Macro-F1 among all models. Overall, FAF-Net achieved the best accuracy, macro-recall, and macro-F1 among the compared backbone models, while maintaining a compact model size.

Table 3. Comparative analysis.

Method	Accuracy	Macro Precision	Macro Recall	Macro F1-Score
VGG16 + Aug	0.5932	0.6080	0.6027	0.5804
ResNet50 + Aug	0.6779	0.6935	0.6722	0.6600
DenseNet121 + Aug	0.6907	0.7057	0.6607	0.6607
InceptionV3 + Aug	0.6525	0.6910	0.6528	0.6473
Baseline	0.6949	0.7107	0.6880	0.6818
Ours (FAF-Net)	0.7458	0.7465	0.7402	0.7291

4.4. Computational Efficiency Analysis

The computational efficiency results are shown in Table 4. Compared with the EfficientNet-B3 baseline, adding the FFT-based frequency branch increased the parameter count from 10.739 M to 10.863 M, and the MACs from 1.930 G to 2.161 G. Adding the auxiliary branch introduced only a very small increase in parameters, from 10.739 M to 10.742 M, without increasing MACs. The complete FAF-Net contained 10.866 M parameters and 2.161 G MACs, indicating that the proposed modules introduced only moderate additional computational overhead.

In the network-forward benchmark, FAF-Net achieved a latency of 16.493 ms/image and a throughput of 60.63 FPS. This result measures the computational cost of the neural network itself when both RGB and frequency-domain input tensors are already prepared. To further evaluate the practical cost introduced by frequency preprocessing, we additionally measured the complete FAF-Net inference pipeline by including GPU-based FFT magnitude preprocessing before the network forward pass. Under this setting, the latency increased to 18.926 ms/image, corresponding to 52.84 FPS. This indicates that FFT preprocessing introduces an additional cost of approximately 2.433 ms/image, but the complete inference pipeline still maintains real-time processing capability.

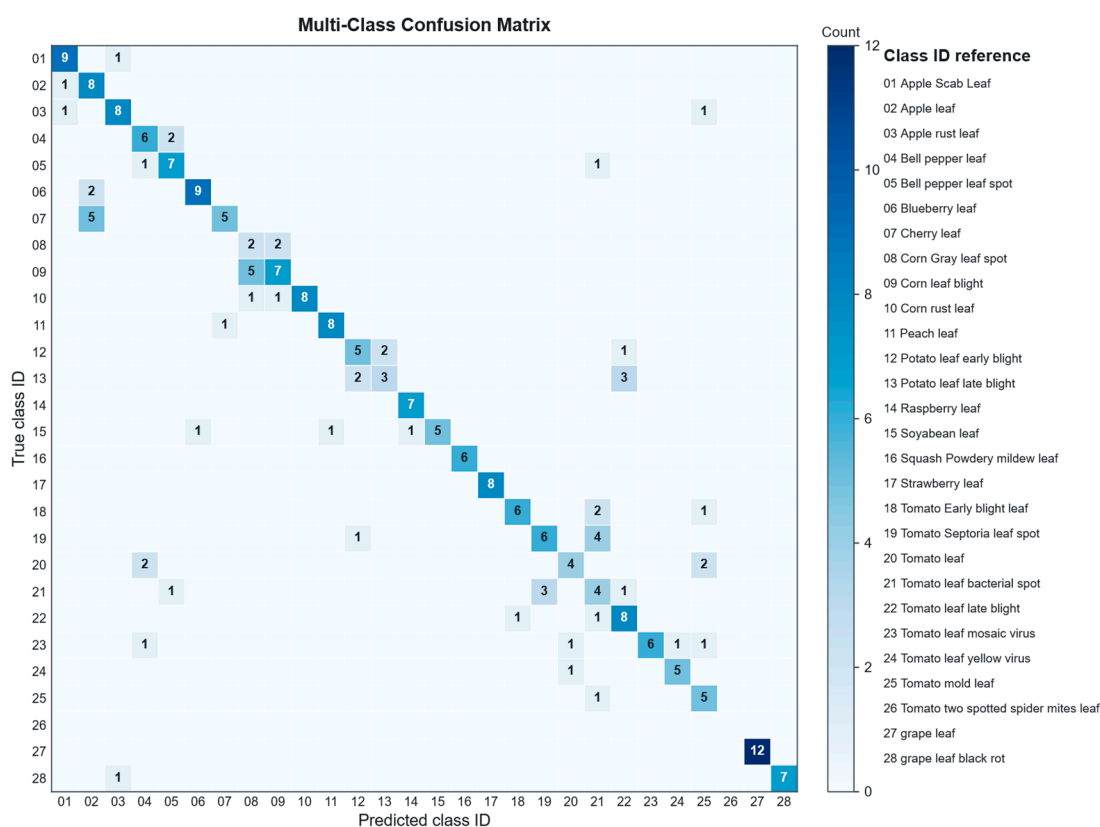
The model size increased only from 41.51 MB for the EfficientNet-B3 baseline to 42.00 MB for FAF-Net, and the peak GPU memory increased from 79.05 MB to 80.59 MB. These results suggest that the FFT-based frequency branch and auxiliary supervision improve classification performance while largely preserving the lightweight property of the EfficientNet-B3 backbone.

Table 4. Computational efficiency of different model variants under the network-forward benchmark.

Model	Params (M)	MACs (G)	FLOPs (G)	Size (MB)	Network Latency (ms/img)	Network FPS	Peak Memory (MB)
Baseline	10.739	1.930	3.860	41.51	14.171	70.57	79.05
Baseline + FFT	10.863	2.161	4.322	41.99	15.210	65.75	80.24
Baseline + Aux	10.742	1.930	3.860	41.52	14.996	66.68	79.77
FAF-Net	10.866	2.161	4.322	42.00	16.493	60.63	80.25

4.5. Confusion Matrix

Figure 8 presents the confusion matrix of FAF-Net on the PlantDoc test set. The rows represent the ground-truth labels, and the columns represent the predicted labels. Most predictions are distributed along the main diagonal, showing that the model can correctly identify most plant leaf categories.

**Figure 8.** Multi-class confusion matrix on the PlantDoc test set.

Several representative misclassified samples are presented and compared with correctly classified samples from the corresponding predicted classes. As shown in Figure 9, a clear confusion pattern appears among visually similar disease categories. For example, corn leaf blight is often misclassified as corn gray leaf spot. These categories usually share similar symptoms, such as spots and edge discoloration, which makes discrimination difficult, especially under natural-scene images.

Another type of error occurs among visually similar healthy leaf categories. Some cherry leaf samples are predicted as apple leaf. When disease symptoms are absent, the model may rely more on the morphological pattern of leaves.

These representative examples suggest that the main errors of FAF-Net are not caused by a general failure of recognition, but are mainly associated with visual similarity

among fine-grained categories in natural-scene images. The model can distinguish the overall healthy/diseased status relatively well, but the discrimination of specific disease types or healthy crop categories remains limited by inter-class similarity, background interference, and image-level supervision. This finding also explains why the binary healthy/diseased confusion matrix showed much better performance than the 27-class classification result. It further suggests that lesion localization, weakly supervised attention, or segmentation-assisted information may be useful for improving fine-grained disease classification in future work.

Representative Misclassified Examples from Visually Similar Classes

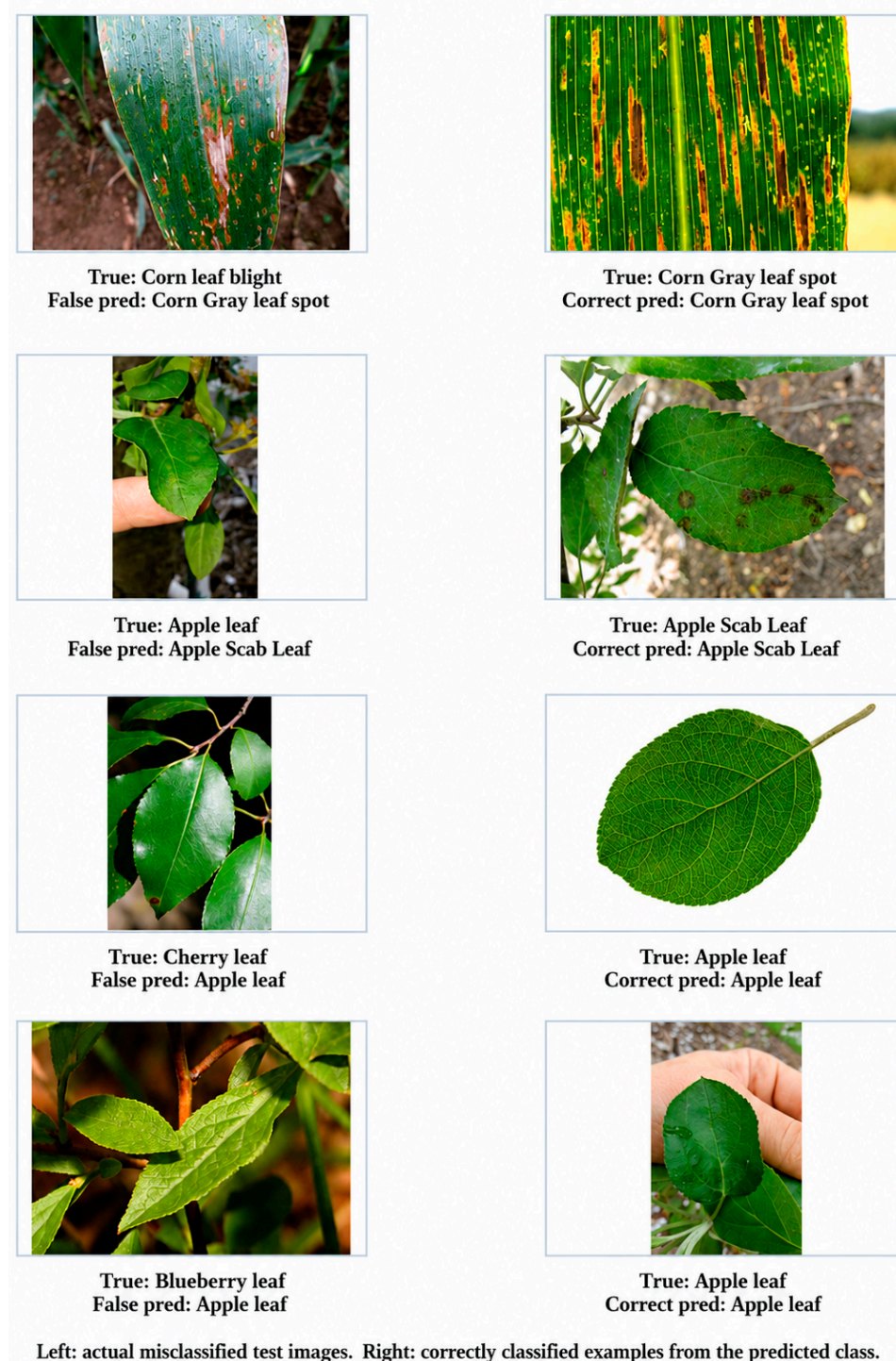


Figure 9. Representative misclassified examples from visually similar classes.

In addition, we construct a binary confusion matrix, shown in Figure 10, by grouping all categories into healthy and diseased leaves. The binary confusion matrix further illustrates that the model is more reliable at the healthy/diseased level than at the category level. Based on the binary confusion matrix, the model achieved an accuracy of 96.19%, a precision of 96.60%, a recall of 97.26%, and an F1-score of 96.93% for healthy/diseased classification. These results indicate that FAF-Net has a strong ability to distinguish the overall health status of leaves. The result supports the motivation for introducing the auxiliary binary supervision branch.

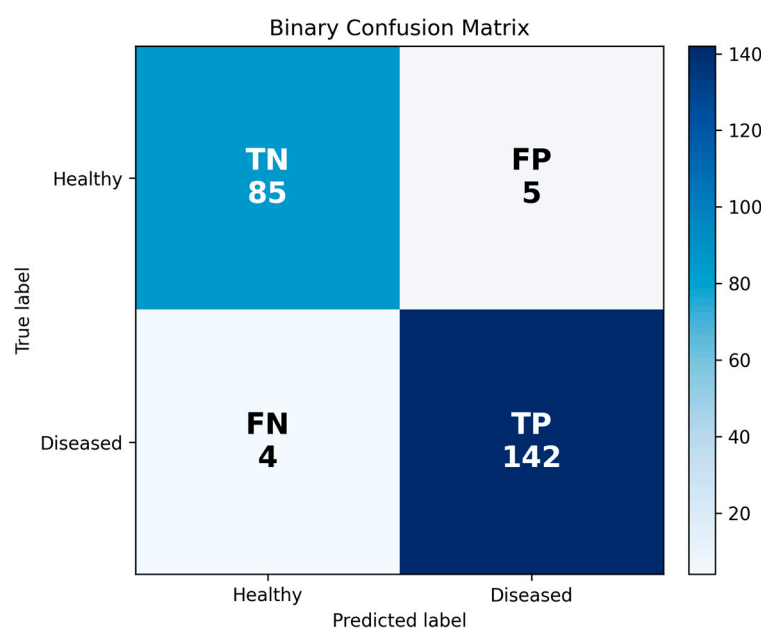


Figure 10. Binary confusion matrix.

5. Discussion

The experimental results show that FAF-Net improves plant disease classification on the natural-scene PlantDoc dataset. On the PlantDoc test set, FAF-Net achieved an accuracy of 74.58%, outperforming the EfficientNet-B3 baseline by 5.09 percentage points. More importantly, the ablation results indicate that the performance gain was not produced by a single component, but by the complementary effects of data augmentation, frequency-domain feature learning, and auxiliary supervision.

Compared with lesion-level localization and segmentation-assisted approaches, FAF-Net works under a lower-annotation setting. Methods such as PlantSeg [30], PDSNets [31], and lesion segmentation and recognition networks [32] use pixel-level masks, disease-spot segmentation, or severity-related annotations to guide the model toward diseased regions. This supervision is effective for lesion localization and severity estimation, but it requires additional manual annotation. FAF-Net uses complete natural-scene images and image-level category labels only. Under this setting, the accuracy increased from 69.49% with EfficientNet-B3 to 74.58% with FAF-Net, indicating that useful disease-related cues can still be strengthened without lesion-level labels. The remaining errors in the confusion matrix also reflect the limitations of this setting. Small lesions, complex background textures, and visually similar symptoms remain difficult to separate without explicit lesion-region supervision.

The frequency branch provides another explanation for the observed improvement. Recent frequency-aware models, including DWTFormer [21], FA-Unet [22], WaveLiteNet [23], and WGA-YOLO [24], have shown that frequency information is useful for

representing lesion boundaries and texture details. In FAF-Net, this idea is implemented as a compact FFT-based branch rather than a large additional backbone. The branch extracts frequency-domain texture information from the Fourier magnitude spectrum and fuses it with EfficientNet-B3 spatial features. The ablation results show that adding the FFT branch after CutMix further improved the classification accuracy, suggesting that frequency-domain cues can complement spatial features when symptoms are small, blurred, or mixed with background patterns. This result supports the use of lightweight frequency complementation for natural-scene image-level disease classification.

The auxiliary supervision branch provides an additional explanation for the improvement from the perspective of training supervision. Existing multi-label or multi-task methods, such as the framework proposed by Yang et al. [26] and the multi-task system developed by Keceli et al. [27], often introduce parallel targets such as plant type, species identity, disease category, or disease severity. These targets enrich the diagnostic output, but they usually require additional labels or a more complex prediction structure. In FAF-Net, the auxiliary task is generated from the original PlantDoc category labels by grouping the classes into healthy and diseased leaves. The main classifier learns the 27-category classification task, while the auxiliary classifier provides coarse-grained disease-status supervision during training. The binary confusion matrix shows much higher performance at the healthy/diseased level than at the 27-category level, suggesting that disease-status information is a stable supervisory signal for this dataset. This label-derived auxiliary task helps guide the shared representation toward disease-relevant features without increasing the annotation burden.

Our confusion matrix provides a more detailed view of where the model succeeds and where it still fails. Most predictions are concentrated along the diagonal, indicating that FAF-Net can correctly classify a large proportion of PlantDoc images. However, the remaining errors are mainly concentrated among visually similar disease categories and morphologically similar healthy categories.

The mixed-dataset experiment further extends the analysis from class-level errors to dataset-level effects. To further analyze the influence caused by datasets, we mixed the PlantDoc and PlantVillage datasets with certain ratios, since PlantDoc and PlantVillage contain partially overlapping crop and disease categories, but they differ substantially in image acquisition conditions and scene complexity. We applied the proposed FAF-Net to train models on datasets with different PlantDoc/PlantVillage ratios, comparing their performance. As shown in Figure 11, as the proportion of PlantDoc increases, the model performance continuously drops.

Specifically, when the PlantDoc dataset proportion is 25%, the model achieved an accuracy of 92.80%. When the ratio of PlantDoc increased to 90%, the accuracy decreased to 76.27%. The lower panel further shows that the absolute performance decline becomes larger when the dataset shifts from a PlantVillage-dominant setting to a PlantDoc-dominant setting. Because all four metrics follow the same downward trend, the decline is unlikely to be a random fluctuation of a single evaluation index.

PlantDoc mainly consists of natural-scene or web-collected images with complex backgrounds, non-uniform illumination, multiple-leaf occlusion, scale variations, different shooting angles, and unstable image quality. PlantDoc also has a relatively small sample size and an imbalanced class distribution, with some categories containing very limited numbers of samples. Under this condition, it is difficult for the model to learn stable fine-grained disease representations for small-sample categories.

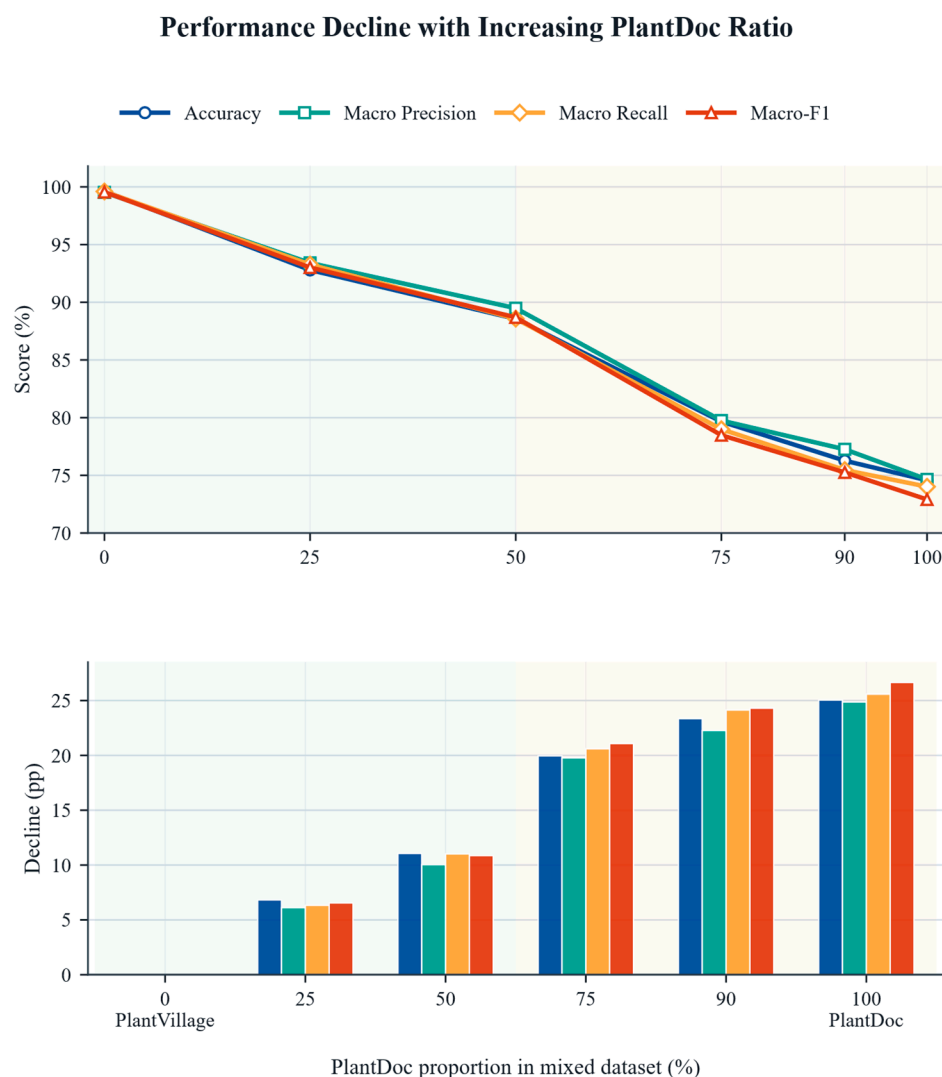


Figure 11. Performance decline with increasing PlantDoc ratio.

In addition, manual inspection shows that some PlantDoc images contain unclear leaf regions, small lesion areas, ambiguous class boundaries, or visually similar symptoms. These issues further increase confusion among fine-grained disease categories, especially among classes with similar symptoms, such as blight, spot, rust, and scab. Therefore, the results in Figure 11 indicate that FAF-Net can alleviate the difficulty of natural-scene classification to some extent, but it cannot completely solve the performance degradation caused by domain shift, background interference, and inter-class similarity.

Several limitations remain in this study. First, the experiments are mainly based on the public PlantDoc dataset, which has a limited sample size and an imbalanced class distribution; therefore, the current results are not sufficient to support strong claims about broad generalization across diverse field environments. Second, the current model uses only image-level labels and does not explicitly localize lesion regions, so background textures, leaf morphology, and imaging conditions may still interfere with classification results. Third, although the FFT-based branch enhances texture representation, it operates on the frequency spectrum of the whole image. As a result, high-frequency responses may also be produced by background elements such as branches, soil, shadow edges, or weed textures. These observations explain why FAF-Net improves PlantDoc classification performance but cannot completely overcome the degradation caused by domain shift, background interference, class imbalance, and inter-class similarity.

Future work can be conducted from both the data and model perspectives. From the data perspective, more field-collected and carefully annotated natural-scene disease images are needed to evaluate the model under realistic agricultural conditions. Diseased leaves from PlantVillage can be pasted onto natural backgrounds to build a larger, more balanced and realistic extended dataset. From the model perspective, leaf segmentation, weakly supervised lesion localization, or lesion-aware attention mechanisms could be introduced to reduce background interference and enhance lesion-region representation. Future studies should evaluate FAF-Net on additional field-collected datasets and independent test sets to more comprehensively assess its stability and generalization ability under diverse natural-scene conditions.

Overall, the mixed-dataset experiment and confusion matrix analysis indicate that the natural-scene complexity and data-quality issues of PlantDoc are important factors limiting further performance improvement. FAF-Net improves classification performance on PlantDoc without relying on lesion-level annotations, but higher-quality datasets and stronger modeling methods are still needed to further improve the reliability of natural-scene plant disease recognition.

6. Conclusions

This study proposed FAF-Net, a frequency-aware fusion network with auxiliary supervision for plant disease classification in natural scenes. The proposed framework uses EfficientNet-B3 as the RGB spatial backbone, integrates an FFT-based frequency branch through learnable weighted fusion, and introduces a healthy/diseased auxiliary supervision task during training. CutMix augmentation was further adopted to improve the robustness of the model against localized symptoms and background interference.

Experiments on the PlantDoc dataset demonstrated that FAF-Net improved the classification performance compared with the EfficientNet-B3 baseline. The classification accuracy increased from 69.49% to 74.58%, and the macro-F1 score increased from 68.18% to 72.91%. Ablation experiments showed that CutMix augmentation, frequency-domain feature fusion, and auxiliary supervision provided complementary improvements. In addition, the healthy/diseased analysis indicated that the model could capture coarse-grained disease-status information more reliably than fine-grained disease categories.

These results suggest that frequency-aware feature fusion and coarse-grained auxiliary supervision are useful for improving classification on the PlantDoc natural-scene benchmark. Nevertheless, fine-grained recognition on PlantDoc remains challenging because of limited sample size, class imbalance, complex backgrounds, and visual similarity among disease categories. Future work will focus on larger field-collected datasets, lesion-aware localization or segmentation, and cross-dataset validation to further improve the robustness and generalization ability of natural-scene plant disease recognition models.

Author Contributions: Conceptualization, C.Q.Z.; methodology, C.Q.Z.; software, C.Q.Z.; validation, C.Q.Z.; formal analysis, C.Q.Z.; investigation, C.Q.Z.; resources, C.Q.Z.; data curation, C.Q.Z.; writing—original draft preparation, C.Q.Z.; writing—review and editing, C.Q.Z. and F.L.S.; visualization, C.Q.Z.; supervision, F.L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study are publicly available from the corresponding sources cited in the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Food and Agriculture Organization of the United Nations. The hidden health crisis: How plant diseases threaten global food security. FAO, 2025. Available online: <https://www.fao.org/one-health/highlights/how-plant-diseases-threaten-global-food-security/en>. (accessed on 13 June 2026).
- Barbedo, J.G.A. Digital image processing techniques for detecting, quantifying and classifying plant diseases. *SpringerPlus* **2013**, *2*, 660. <https://doi.org/10.1186/2193-1801-2-660>.
- Bock, C.H.; Poole, G.H.; Parker, P.E.; Gottwald, T.R. Plant disease severity estimated visually, by digital photography and image analysis, and by hyperspectral imaging. *Crit. Rev. Plant Sci.* **2010**, *29*, 59–107. <https://doi.org/10.1080/07352681003617285>.
- Padol, P.B.; Yadav, A.A. SVM classifier based grape leaf disease detection. In Proceedings of the Conference on Advances in Signal Processing (CASP), Pune, India, 9–11 June 2016; pp. 175–179. <https://doi.org/10.1109/CASP.2016.7746160>.
- Singh, V.; Misra, A.K. Detection of plant leaf diseases using image segmentation and soft computing techniques. *Inf. Process. Agric.* **2017**, *4*, 41–49. <https://doi.org/10.1016/j.inpa.2016.10.005>.
- Pydipati, R.; Burks, T.F.; Lee, W.S. Identification of citrus disease using color texture features and discriminant analysis. *Comput. Electron. Agric.* **2006**, *52*, 49–59. <https://doi.org/10.1016/j.compag.2006.01.004>.
- Liu, J.; Wang, X. Plant diseases and pests detection based on deep learning: A review. *Plant Methods* **2021**, *17*, 22. <https://doi.org/10.1186/s13007-021-00722-9>.
- LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. <https://doi.org/10.1038/nature14539>.
- Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; Chen, T. Recent advances in convolutional neural networks. *Pattern Recognit.* **2018**, *77*, 354–377. <https://doi.org/10.1016/j.patcog.2017.10.013>.
- Wang, S.; Xu, D.; Liang, H.; Bai, Y.; Li, X.; Zhou, J.; Su, C.; Wei, W. Advances in deep learning applications for plant disease and pest detection: A review. *Remote Sens.* **2025**, *17*, 698. <https://doi.org/10.3390/rs17040698>.
- Mohanty, S.P.; Hughes, D.P.; Salathé, M. Using deep learning for image-based plant disease detection. *Front. Plant Sci.* **2016**, *7*, 1419. <https://doi.org/10.3389/fpls.2016.01419>.
- Sladojevic, S.; Arsenovic, M.; Anderla, A.; Culibrk, D.; Stefanovic, D. Deep neural networks based recognition of plant diseases by leaf image classification. *Comput. Intell. Neurosci.* **2016**, *2016*, 3289801. <https://doi.org/10.1155/2016/3289801>.
- Ferentinos, K.P. Deep learning models for plant disease detection and diagnosis. *Comput. Electron. Agric.* **2018**, *145*, 311–318. <https://doi.org/10.1016/j.compag.2018.01.009>.
- Too, E.C.; Yujian, L.; Njuki, S.; Yingchun, L. A comparative study of fine-tuning deep learning models for plant disease identification. *Comput. Electron. Agric.* **2019**, *161*, 272–279. <https://doi.org/10.1016/j.compag.2018.03.032>.
- Barbedo, J.G.A. Factors influencing the use of deep learning for plant disease recognition. *Biosyst. Eng.* **2018**, *172*, 84–91. <https://doi.org/10.1016/j.biosystemseng.2018.05.013>.
- Hughes, D.P.; Salathé, M. An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv* **2015**, arXiv:1511.08060. <https://doi.org/10.48550/arXiv.1511.08060>.
- Singh, D.; Jain, N.; Jain, P.; Kayal, P.; Kumawat, S.; Batra, N. PlantDoc: A dataset for visual plant disease detection. In Proceedings of the 7th ACM IKDD CoDS and 25th COMAD, Hyderabad, India, 5–7 January 2020; pp. 249–253. <https://doi.org/10.1145/3371158.3371196>.
- Tan, M.; Le, Q.V. EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; PMLR 97; pp. 6105–6114. Available online: <https://proceedings.mlr.press/v97/tan19a.html> (accessed on 13 June 2026).
- Rahaman, N.; Baratin, A.; Arpit, D.; Draxler, F.; Lin, M.; Hamprecht, F.A.; Bengio, Y.; Courville, A. On the spectral bias of neural networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; PMLR 97; pp. 5301–5310. Available online: <https://proceedings.mlr.press/v97/rahaman19a.html> (accessed on 13 June 2026).
- Li, H.; Shi, H.; Du, A.; Mao, Y.; Fan, K.; Wang, Y.; Shen, Y.; Wang, S.; Xu, X.; Tian, L.; Wang, H.; Ding, Z. Symptom recognition of disease and insect damage based on Mask R-CNN, wavelet transform, and F-RNet. *Front. Plant Sci.* **2022**, *13*, 922797. <https://doi.org/10.3389/fpls.2022.922797>.
- Xiang, Y.; Gao, S.; Li, X.; Li, S. DWTFormer: A frequency-spatial features fusion model for tomato leaf disease identification. *Plant Methods* **2025**, *21*, 33. <https://doi.org/10.1186/s13007-025-01349-w>.
- Li, X.; Wu, W.; Zhu, F.; Guan, S.; Zhang, W.; Li, Z. FA-Unet: A deep learning method with fusion of frequency domain features for fruit leaf disease identification. *Horticulturae* **2025**, *11*, 783. <https://doi.org/10.3390/horticulturae11070783>.

23. Yang, J.; Xu, G.; Yang, M.; Lin, Z. Lightweight wavelet-CNN tea leaf disease detection. *PLoS ONE* **2025**, *20*, e0323322. <https://doi.org/10.1371/journal.pone.0323322>.
24. Zhao, J.; Liang, Y.; Wei, G.; Han, B.; Li, J.; Yu, C.; Zeng, J.; He, S. A wavelet-based frequency-domain approach for accurate multi-crop disease detection. *Sci. Rep.* **2026**, *16*, 7099. <https://doi.org/10.1038/s41598-026-38476-w>.
25. Zhang, L.; Yu, M.; Chen, T.; Shi, Z.; Bao, C.; Ma, K. Auxiliary training: Towards accurate and robust models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 372–381. <https://doi.org/10.1109/CVPR42600.2020.00045>.
26. Yang, B.; Li, M.; Li, F.; Wang, Y.; Liang, Q.; Zhao, R.; Li, C.; Wang, J. A novel plant type, leaf disease and severity identification framework using CNN and transformer with multi-label method. *Sci. Rep.* **2024**, *14*, 11664. <https://doi.org/10.1038/s41598-024-62452-x>.
27. Keceli, A.S.; Kaya, A.; Catal, C.; Tekinerdogan, B. Deep learning-based multi-task prediction system for plant disease and species detection. *Ecol. Inform.* **2022**, *69*, 101679. <https://doi.org/10.1016/j.ecoinf.2022.101679>.
28. Salman, Z.S.Z.; Muhammad, A.M.A.; Han, D.H. Plant disease classification in the wild using vision transformers and mixture of experts. *Front. Plant Sci.* **2025**, *16*, 1522985. <https://doi.org/10.3389/fpls.2025.1522985>.
29. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. CutMix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6023–6032. <https://doi.org/10.1109/ICCV.2019.00612>.
30. Wei, T.; Chen, Z.; Yu, X.; Chapman, S.; Melloy, P.; Huang, Z. A Large-Scale In-the-wild Dataset for Plant Disease Segmentation. *Sci. Data* **2026**, *13*, 205. <https://doi.org/10.1038/s41597-025-06513-4>.
31. Zhao, C.; Li, C.; Wang, X.; Wu, X.; Du, Y.; Chai, H.; Cai, T.; Xiang, H.; Jiao, Y. Plant Disease Segmentation Networks for Fast Automatic Severity Estimation Under Natural Field Scenarios. *Agriculture* **2025**, *15*, 583. <https://doi.org/10.3390/agriculture15060583>.
32. Yang, T.; Wang, Y.; Lian, J. Plant Diseased Lesion Image Segmentation and Recognition Based on Improved Multi-Scale Attention Net. *Appl. Sci.* **2024**, *14*, 1716. <https://doi.org/10.3390/app14051716>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.