

Article

Tomato Leaf Disease Identification via Information-Theoretic Entropy Attention and Hierarchical Feature Alignment

Zhiyi Sun ^{1,*}, Shengying Yang ^{1,*} , Jianfeng Wu ^{2,*} and Boyang Feng ³

¹ School of Computer Science and Technology, Zhejiang University of Science and Technology, Hangzhou 310023, China; 1230204088@zust.edu.cn

² College of Information Science & Technology, Zhejiang Shuren University, Hangzhou 310015, China

³ The University of Melbourne, Melbourne 3000, Australia

* Correspondence: syyang@zust.edu.cn (S.Y.); wujianfeng@zjsru.edu.cn (J.W.)

Abstract

Tomato, as a globally vital economic crop, relies heavily on accurate disease recognition to safeguard food security. However, tomato leaf disease identification constitutes a classic fine-grained visual classification task characterized by minimal inter-class variance, spatially sparse lesion features, and complex background interference. These challenges hinder conventional deep learning models from precisely localizing critical discriminative regions. In response to the aforementioned challenges, we introduce EA-HFA, an innovative framework based on deep neural networks that synergistically integrates an Entropy Attention mechanism alongside a Hierarchical Feature Alignment component. Specifically, the Entropy Attention module leverages information-theoretic entropy to quantify pixel-wise predictive uncertainty, adaptively selecting high-confidence pixels to automatically focus the network on sparse yet highly discriminative lesion features. Concurrently, the Hierarchical Feature Alignment module imposes KL-divergence constraints on the temperature-scaled probability distributions across adjacent network layers, enforcing cross-scale consistency in the localization of discriminative regions. Evaluations conducted on the PlantVillage and AI Challenger 2018 benchmarks reveal that EA-HFA achieves Top-1 accuracies of 99.29% and 97.82%, respectively, yielding performance comparable to established deep learning architectures while maintaining a reasonable computational footprint. Furthermore, qualitative analyses indicate that the model tends to attend to minute lesion-relevant areas, providing a certain level of interpretability for its decision-making process. Thus, EA-HFA holds practical potential as an alternative solution for automated plant disease monitoring in precision farming.

Keywords: tomato leaf disease; entropy attention; hierarchical feature alignment; fine-grained visual classification; deep neural networks; KL divergence; plant disease recognition; interpretability; precision farming



Academic Editor: Maciej Zaborowicz

Received: 1 May 2026

Revised: 24 June 2026

Accepted: 25 June 2026

Published: 29 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

Ranking only behind the potato in terms of global production volume, the tomato is indispensable to worldwide food security and economies. However, tomatoes face significant threats from diverse pathological conditions which severely compromise both yield and quality. Consequently, the accurate identification of disease symptoms is critical for ensuring sustainable tomato production [1]. Traditional methods for crop disease recognition rely heavily on expert empirical judgment, which is not only labor-intensive

but also prone to subjectivity [2,3]. Benefiting from advancements in computer vision, deep neural networks have gained significant traction within agricultural pathology. Deep learning models possess the capability to automatically learn and extract discriminative features directly from raw image pixels, a capacity that serves as the cornerstone for efficient and accurate crop disease diagnosis [4].

As illustrated in Figure 1, different categories of plant leaf pathologies, such as Early and Late Blight, exhibit high similarity in their macroscopic morphology, often manifesting generically as wilting or spotting. The discriminative differences are confined to subtle local features, such as the texture of lesion margins or spore distribution patterns [3]. Conversely, environmental variations, including changes in illumination and viewing angles, can induce significant intra-class variability. This characteristic—large intra-class variance coupled with small inter-class differences—defines plant leaf disease recognition as a classic fine-grained classification task [5]. Existing studies [6,7] indicate that in leaf disease imagery, the discriminative signals truly determining the classification occupy only a small fraction of the image, while the majority of the region consists of common background or non-discriminative textures. Under such fine-grained conditions, it is imperative for models to precisely localize these local discriminative regions to achieve robust performance.

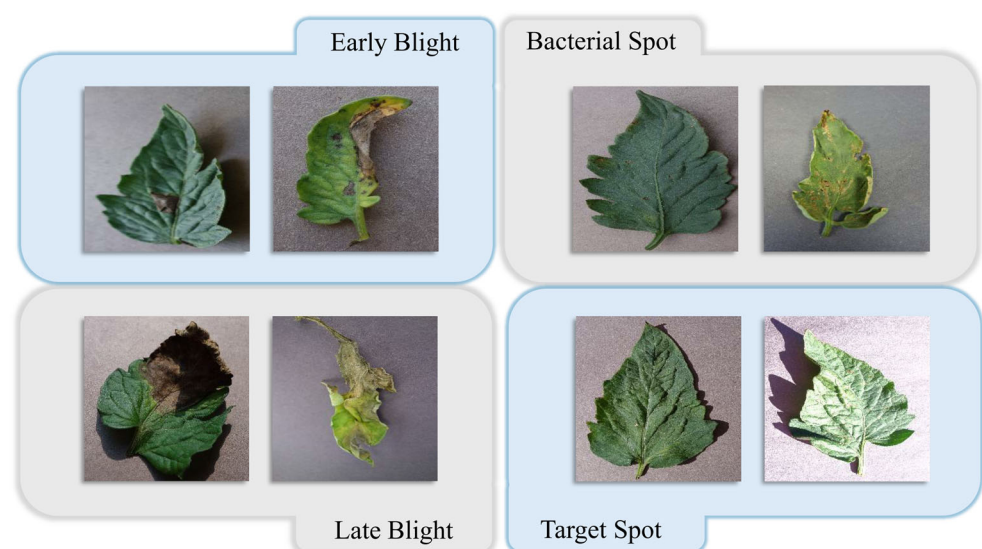


Figure 1. Example images of four tomato leaf diseases with high inter-class similarity and intra-class variation.

Traditional deep learning models, including ResNet, have demonstrated remarkable success in feature extraction. However, they typically rely on mechanisms such as global pooling, which aggregates information across the entire image [8]. While effective for general classification, this holistic approach may inadvertently incorporate background noise and non-discriminative regions when discriminative features are spatially sparse. Consequently, although these models can achieve high overall accuracy, their decision-making processes often lack interpretability, as they might rely on broader contextual cues rather than exclusively focusing on subtle lesion characteristics. As illustrated by the Grad-CAM visualizations in Figure 2, existing models sometimes focus heavily on low-level textural details or surrounding environments in shallow layers, while in deeper layers, they may struggle to maintain consistent spatial alignment strictly with the target disease symptoms, occasionally attending to irrelevant background areas.

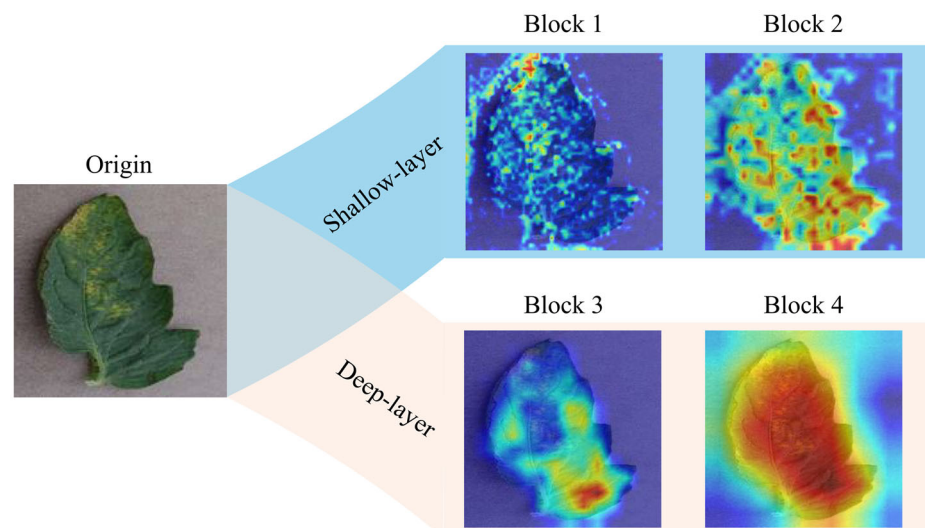


Figure 2. Grad-CAM visualizations of a ResNet50 model on tomato leaf disease images. Shallow-layer features predominantly activate on background textures, while deep-layer features exhibit spatial misalignment with precise lesion boundaries. The heatmap overlays the original image, where warmer colors (red) indicate higher activation, and cooler colors (blue) indicate lower activation.

The human visual system selectively allocates attention to salient regions and is highly sensitive to distinct features. Inspired by the brain’s ability to efficiently compress and represent raw redundant information, researchers have established the theory of sparse coding [9,10]. Building on this inspiration, attention mechanisms have emerged as a fundamental component within modern neural networks [11]. Drawing inspiration from this paradigm, we posit that to address the spatially sparse distribution of discriminative signals in tomato disease imagery, models must be equipped with the intrinsic capability to automatically localize and focus on these critical regions, effectively mimicking the representational efficiency of the brain.

To tackle the challenges of spatial sparsity in discriminative features and the difficulty existing models face in precisely localizing subtle lesions within the fine-grained classification of tomato leaf diseases, this paper proposes an Information-Theoretic Entropy Attention Mechanism. Entropy quantifies the degree of disorder or unpredictability of a random variable [12]. Our mechanism leverages information entropy to measure pixel-wise predictive uncertainty. Theoretically, low entropy values signify high model confidence in the classification of a specific region, which typically corresponds to key discriminative features such as characteristic lesion morphologies. By employing the Straight-Through Estimator (STE) to binarize and select the Top-K low-entropy points, the model effectively concentrates its computational resources on these sparse, high-value regions. Furthermore, to address the issues whereby shallow networks are prone to textural noise interference and deep networks may drift away from lesion regions, we designed a Hierarchical Feature Alignment module. This module imposes a Kullback–Leibler (KL) divergence constraint on the temperature-scaled probability distributions of adjacent layers, encouraging the discriminative regions identified at different scales to maintain consistency. This strategy aims to mitigate erroneous selections induced by noise, helping to preserve feature localization across the network hierarchy.

The main contributions of this study are summarized as follows:

1. We propose an information entropy-based discriminative feature selection attention module. Addressing the challenge of sparsely distributed disease features, this module utilizes information entropy to quantify pixel-wise uncertainty and adaptively focuses on low-entropy, highly discriminative lesion regions. It effectively suppresses

- the dilution of critical signals by complex background noise, solving the problem that traditional models struggle to precisely localize subtle disease spots.
2. We construct a Hierarchical Feature Alignment self-distillation module. By introducing a KL divergence constraint on the probability distributions of adjacent layers, this module mitigates the inconsistency between shallow and deep networks in terms of semantics and spatial localization.
 3. The proposed EA-HFA achieves Top-1 accuracies of 99.29% and 97.82% on the PlantVillage and AI Challenger datasets, respectively, yielding performance comparable to mainstream deep learning architectures while maintaining a reasonable computational footprint. Furthermore, qualitative analysis suggests that this method tends to concentrate on subtle pathological features in disease images rather than attending to irrelevant backgrounds, offering a practical and interpretable alternative for future monitoring tools in smart agriculture.

2. Related Work

Early approaches to computer-based plant leaf disease recognition primarily relied on multi-stage traditional methods, including techniques like K-means, Support Vector Machines, and the Otsu method [13–15]. However, these methods suffered from suboptimal performance due to their reliance on manually engineered descriptors and their inability to extract sufficiently representative characteristics [5,16]. Following the introduction of AlexNet [17], Deep Convolutional Neural Networks have been extensively applied to plant disease recognition. Architectures such as VGG [18], ResNet [19], and DenseNet [20] can automatically learn features directly from data, achieving significantly higher accuracy [21].

In the specific domain of tomato disease recognition, deep learning has rapidly become the mainstream paradigm. Brahimi et al. [22] proposed a CNN-based classification method for tomato diseases, achieving a classification accuracy of 99.185% on large-scale datasets via transfer learning. Sardogan et al. [23] developed a fused framework integrating CNNs with Learning Vector Quantization; by extracting color features and feeding them into an LVQ classifier, they attained an average accuracy of 86% across five tomato leaf categories. Furthermore, Rangarajan et al. [24] and Maeda-Gutiérrez et al. [25] conducted systematic comparisons of various CNN architectures for tomato disease recognition. These studies collectively validate the superior performance of CNN models, highlighting the superior efficacy of DL-based approaches compared to conventional techniques.

In recent years, researchers have sought to further enhance model performance through more powerful feature extraction and fusion mechanisms. Borhani et al. [21] proposed a hybrid Vision Transformer and CNN-Transformer architecture for real-time plant disease classification, demonstrating that attention mechanisms can substantially improve accuracy. Sanida et al. [26] developed a hybrid CNN model fusing VGG and Inception modules, achieving a test accuracy of 99.17% on the PlantVillage dataset. Attallah et al. [27] proposed a method that fuses deep features from three distinct CNNs (ResNet-18, ShuffleNet, and MobileNet) and employs a hybrid feature selection strategy, realizing a remarkable 99.92% accuracy in tomato disease classification. Addressing the issue of insufficient generalization to Out-of-Distribution samples, Li et al. [16] proposed MLFAnet. This method utilizes the Discrete Cosine Transform to extract multi-low-frequency components from feature maps and designs a corresponding attention mechanism to effectively suppress high-frequency perturbations. Sun et al. [28] introduced Eff-Swin, a dual-stream network that fuses an improved EfficientNetV2 with a Swin Transformer, achieving 99.70% accuracy on a fused dataset. Additionally, Chen et al. [29] proposed a hybrid model combining Transformers and CNNs, leveraging Generative Adversarial Networks for data augmentation to achieve high-precision, lightweight detection of tomato leaf diseases.

Research on automatically focusing on discriminative regions has also advanced significantly within the field of fine-grained recognition [6]. Ji et al. [30] proposed ACNet, which integrates convolutional kernels along the boundary nodes of a fixed-height binary tree and employs a soft routing decision mechanism to achieve coarse-to-fine hierarchical feature learning and discriminative region localization. Lu et al. [31] introduced a framework for learning local structural information based on multi-directional filters, effectively extracting local image features without requiring bounding boxes or part annotations. Chou et al. [32] proposed the HERBS method, which employs dual components for background suppression and high-temperature refinement to inhibit background noise and focus on fine-grained features. These methodologies provide valuable insights and a strong foundation for advancing tomato disease recognition.

Distinctions from Existing Methods. There are fundamental differences between our proposed modules and existing methods. Traditional attention mechanisms typically learn a soft weight map multiplied onto features, which often retains a substantial amount of background noise. In contrast, our Entropy Attention directly discards non-discriminative regions through hard selection, facilitating more effective sparse feature focusing. Furthermore, while traditional uncertainty estimation is generally used to evaluate overall model confidence, our approach spatializes uncertainty, transforming it into an “attention map” that guides network training towards lesion localization. Unlike standard knowledge distillation that transfers global class probabilities, our Hierarchical Feature Alignment functions as a spatial self-distillation mechanism. It distills temperature-scaled probability distributions to encourage consistency between shallow and deep layers regarding “where to look.” Moreover, while traditional feature alignment primarily addresses resolution and semantic fusion, our Hierarchical Feature Alignment explicitly constrains cross-layer attention consistency without altering the inherent semantics of the features.

3. Materials and Methods

3.1. Proposed Method

3.1.1. EA-HFA Framework Overview

As illustrated in Figure 3, the proposed EA-HFA is a synergistic framework composed of two core modules: Entropy Attention (EA) and Hierarchical Feature Alignment (HFA). This architecture is specifically designed to address the challenge of sparsely distributed discriminative features in tomato leaf disease images. By leveraging these modules, the framework enables the model to automatically localize critical lesion regions across multiple scales, even when such features are sparse and subtle within the input image.

After the raw images are fed into the network, they first pass through the backbone network. To this end, the backbone of the proposed EA-HFA extracts feature maps at multiple hierarchical levels, each corresponding to distinct spatial resolutions and semantic hierarchies. Subsequently, the multi-level feature maps are fed into our EA module, where the generated entropy maps are mainly used for low-entropy region selection to guide the model’s focus. Concurrently, our HFA module imposes consistency constraints on the temperature-scaled class probability distributions across adjacent layers. Finally, the collaboratively optimized features are fed into a Graph Convolutional Network (GCN) for the final classification.

3.1.2. Entropy Attention

As illustrated in Figure 3, the proposed EA module operates on feature maps at every hierarchical level, serving as the core component responsible for distinguishing subtle local feature variations across multi-scale representations.

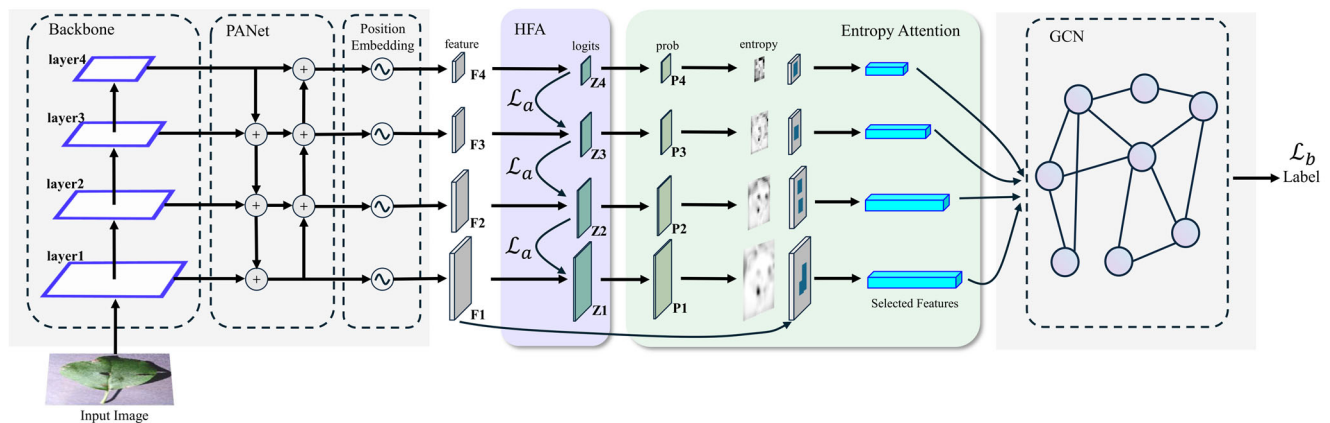


Figure 3. Overview of the proposed EA-HFA framework for tomato leaf disease identification. The architecture consists of four main components: (1) a backbone network (e.g., ResNet) that extracts multi-scale feature maps from input images; (2) a PANet-based feature pyramid module with position embedding to enhance spatial context; (3) the core EA-HFA module, which includes an EA branch that generates pixel-wise uncertainty maps and selects high-confidence discriminative features, and a HFA branch that imposes KL-divergence constraints across adjacent layers to ensure cross-scale consistency; and (4) a Graph Convolutional Network classifier that leverages structured relationships among selected features for final prediction.

In tomato disease imagery, the lesion regions that truly determine the category—such as small disease spots—are localized fine-grained features. These regions typically occupy only a minor fraction of the entire image and often exhibit visual appearances highly similar to non-discriminative areas, such as healthy tissue. In traditional CNNs, features from a vast number of these non-discriminative regions are inadvertently incorporated into the classification decision process, thereby diluting the critical discriminative signals.

Given the sparsity of discriminative signals, we require a mechanism to quantify the uncertainty of each pixel—specifically, through entropy. From the perspective of information theory, as shown in Equation (1), the discriminability of a feature point at (i, j) is equivalent to the amount of information it provides regarding the class label, defined as the mutual information $I(X; Y_{i,j})$.

Since the prior entropy of the class labels, $H(X)$, is constant, maximizing mutual information is mathematically equivalent to minimizing the conditional entropy $H(X|Y_{i,j})$. Consequently, entropy serves as an exact measure of predictive uncertainty: lower entropy indicates higher classification confidence by the model at that specific point. While low-entropy regions may occasionally originate from leaf edges or stable background textures, they are more likely, in most cases, to contain high-confidence discriminative information, such as characteristic lesion morphologies.

$$I(X; Y_{i,j}) = H(X) - H(X|Y_{i,j}) \quad (1)$$

Here, X denotes the ground truth class label of the image, and $Y_{i,j}$ represents the feature representation at spatial location (i, j) . Given a specific feature value $Y_{i,j} = y_{i,j}$, the conditional entropy is defined as shown in Equation (2).

$$H(X|Y_{i,j} = y_{i,j}) = - \sum_{c=1}^{C_k} P(X = c | Y_{i,j} = y_{i,j}) \log P(X = c | Y_{i,j} = y_{i,j}) \quad (2)$$

Here, $P(X = c | Y_{i,j} = y_{i,j})$ represents the true conditional probability distribution. In practical model implementation, by minimizing the cross-entropy loss, we drive the classifier's output probabilities, denoted as $P_l[i, j, c]$, to approximate the true posterior distribu-

tion. Consequently, the entropy map $E_l[i, j]$ derived from the converged model serves as an effective estimator of the theoretical conditional entropy $H(X|Y_{i,j})$.

As demonstrated in Equation (3), even in the presence of minor estimation deviations, the relative ordinal relationship of entropy values between any two feature points (e.g., point A and point B) remains stable.

$$E_l^A \ll E_l^B \Rightarrow H(X|Y_A = y_A) < H(X|Y_B = y_B) \quad (3)$$

In summary, we can conclude that lower entropy at a feature point generally corresponds to stronger discriminability. To prevent the network from being distracted by non-lesion low-entropy regions, cross-layer semantic guidance is required, which motivates our subsequent HFA module. Guided by this insight, the Entropy Attention module employs information entropy as a quantitative metric for discriminability. By automatically selecting low-entropy points, the module achieves effective focusing on lesion regions, effectively filtering out non-discriminative areas.

Specifically, for the l -th layer, let $F_l \in R^{C_l \times H_l \times W_l}$ denote the feature map after feature aggregation and positional embedding, where H_l and W_l represent the spatial height and width, and C_l denotes the number of channels. As formulated in Equation (4), each layer is equipped with a dedicated 1×1 convolutional classifier. The feature map F_l sequentially passes through a feature transformation convolution, batch normalization (BN), an activation function, and a class projection convolution to generate a pixel-wise score map Z_l :

$$Z_l = W_2^{(l)} \cdot \text{ReLU}\left(\text{BN}\left(W_1^{(l)} \cdot F_l\right)\right) \quad (4)$$

Here, $Z_l \in R^{C_k \times H_l \times W_l}$, where C_k represents the number of classes in the classification task. $W_1^{(l)}$ and $W_2^{(l)}$ correspond to the learnable parameters of the first and second convolutional layers, respectively. As shown in Equations (5) and (6), Z_l is processed by a Softmax function to obtain a pixel-wise probability map P_l , from which the entropy map E_l is subsequently calculated:

$$P_l = \text{Softmax}(Z_l) \quad (5)$$

$$E_l[i, j] = - \sum_{c=1}^{C_k} P_l[i, j, c] \cdot \log(P_l[i, j, c] + \epsilon) \quad (6)$$

We perform a Top-K selection on $E_l[i, j]$ (selecting points with the lowest entropy) and employ the Straight-Through Estimator (STE) to ensure gradient backpropagation. The specific operations are defined as follows:

$$W_l^{soft} = \sigma(-\beta \cdot E_l) \cdot \frac{K}{\sum_{i,j} \sigma(-\beta \cdot E_l[i, j])} \quad (7)$$

$$M_l^{hard}[i, j] = \begin{cases} 1 & \text{if selected} \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

$$M_l = M_l^{hard} + \left(W_l^{soft} - \text{detach}\left(W_l^{soft}\right)\right) \quad (9)$$

In these equations, σ denotes the Sigmoid function. During the forward pass, the mask is determined by the hard selection ($M_l = M_l^{hard}$). The STE strategy not only guarantees gradient flow but also exerts a certain degree of suppression on unselected points. This mechanism aids the Entropy Attention module in automatically and effectively focusing on the most discriminative features.

3.1.3. Hierarchical Feature Alignment

As illustrated in Figure 3, the proposed HFA module encourages consistency between the predictive attention distributions of adjacent layers. Rather than directly aligning the scalar entropy maps, we utilize the temperature-scaled class probability distributions, generated from the pixel-wise score maps of feature maps F_l and F_{l-1} , as our proxy for spatial attention. To further mitigate the impact of abundant irrelevant features, this module introduces a self-distillation mechanism to facilitate effective information flow across hierarchies [33–35]. Specifically, it addresses the potential challenges depicted in Figure 2: shallow networks are prone to interference from textural noise, while deep networks may occasionally lose spatial alignment with the precise lesion regions. By aligning these probability distributions across layers, our approach aims to preserve and refine the model's focus on discriminative signals is preserved and refined throughout the network depth.

As illustrated in Figure 4, both F_l and F_{l-1} are processed by 1×1 convolutional classifiers identical to those used in the Entropy Attention module. This design offers flexibility: the classifiers can be shared with the existing Entropy Attention modules to minimize parameter overhead, or implemented as independent components depending on the specific architectural requirements.

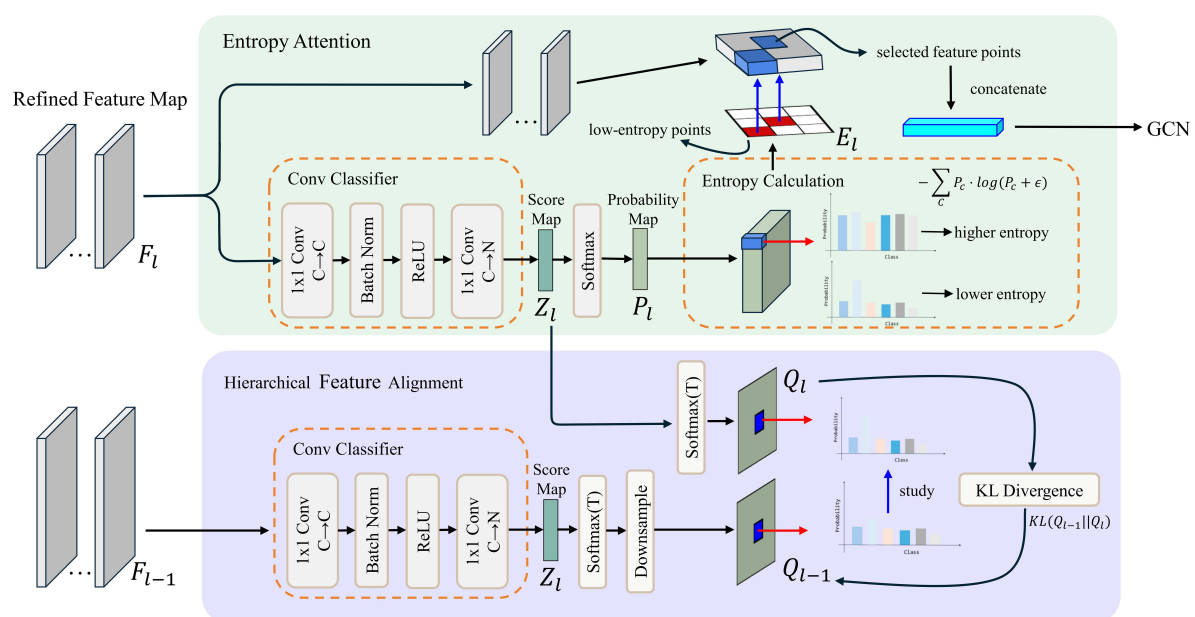


Figure 4. Detailed architecture of the EA and HFA modules. The EA module (**top**) processes a refined feature map (F_l) through a convolutional classifier and entropy calculation to select discriminative feature points by filtering for low-entropy regions. The HFA module (**bottom**) enforces consistency by applying a KL-divergence constraint between the probability distributions of adjacent feature maps (Q_l and Q_{l-1}), aligning the attention focus across different hierarchical levels.

Upon obtaining the score maps Z_l and Z_{l-1} , we introduce a temperature-scaled knowledge distillation strategy to refine these representations. As formulated below, the score maps undergo temperature scaling to soften the probability distributions, facilitating more effective knowledge transfer between layers:

$$Q_l = \text{Softmax}(Z_l/T) \quad (10)$$

$$Q_{l-1} = \text{LogSoftmax}(Z_{l-1}/T) \quad (11)$$

Here, T denotes the temperature parameter, which governs the smoothness of the probability distributions. Initially, a higher temperature is employed to generate smoother distributions, facilitating the transfer of richer structured information (often referred to as “dark knowledge”) between layers. As training progresses, T is gradually annealed, prompting each layer to increasingly focus on the most discriminative features.

Given the discrepancy in spatial resolutions across different hierarchical levels, we apply downsampling to the shallow feature maps to align their dimensions with those of the deeper layers. Finally, the KL divergence is utilized to enforce the shallow layers to learn the temperature-scaled class probability distribution of the deeper layer. The overall loss function for the HFA module is formulated as follows:

$$\mathcal{L}_a = \sum_{l=1}^L T^2 \cdot KL(Q_{l-1} || Q_l) \quad (12)$$

3.1.4. Implementation Details

The Feature Aggregation component employs a Path Aggregation Network (PAN) architecture [36], incorporating both top-down and bottom-up feature pathways. In our implementation, we utilize only the feature maps generated by the final pathway at each hierarchical level. Notably, this design choice underscores the flexibility of the proposed EA-HFA framework, as the aggregation module can be seamlessly substituted with alternative feature fusion mechanisms without compromising the core logic of the subsequent modules.

Prior to entering the Entropy Attention and Hierarchical Feature Alignment modules, the feature maps undergo positional embedding. Theoretically, higher-level embeddings facilitate the learning of object-level relative positions and inter-part spatial relationships, thereby enriching semantic information, while lower-level embeddings encode absolute pixel coordinates to enhance geometric feature representation.

For the ultimate prediction phase, the model aggregates the chosen representative features via concatenation. The subsequent GCN [37] leverages the spatial relationships encoded by the positional embeddings to derive the final classification results.

Following the multi-level entropy-based filtering, the selected sparse feature points from all hierarchical layers are concatenated and fed into a GCN classifier. Instead of treating the massive number of selected points individually, we employ a learnable parametric pooling layer to aggregate them into a compact set of representative graph nodes. Specifically, the total number of points is reduced by a factor of 64 to form the graph nodes. The core of our single-layer GCN lies in its dynamically constructed adjacency matrix, which governs the message passing between nodes. Specifically, we first project the node features into query and key vectors with reduced dimensions ($C/4$). To capture the global activation response of each node and reduce computational overhead, we apply a channel-wise mean pooling to compress these multi-channel vectors into scalar representations. Subsequently, we compute the pairwise scalar differences between all node pairs via a broadcasting mechanism, followed by a hyperbolic tangent (Tanh) activation, to capture the intrinsic relative affinities. This dynamic affinity matrix is then adaptively fused with a learnable prior adjacency matrix (initialized as a slightly perturbed identity matrix). Unlike traditional graph normalization strategies (e.g., symmetric normalized Laplacian), our adjacency matrix is entirely data-driven and adaptively scaled by a learnable parameter α . Traditional Global Average Pooling treats multi-scale features as unordered sets, inevitably discarding their structural relationships. Our GCN addresses this by performing structured message passing (via a 1×1 convolution for feature transformation, Batch Normalization, and matrix multiplication with the adjacency matrix) over the dynamically constructed graph. Finally, another learnable pooling layer aggregates the node features into a global representation, followed by a dropout layer ($p = 0.1$) and a fully connected layer for the

final classification. The detailed forward pass and hyperparameter settings of the GCN classifier are summarized in Algorithm 1.

Algorithm 1. Forward Pass of the Proposed GCN Classifier

Input: Concatenated sparse features $X \in R^{B \times C \times N_{select}}$

Output: Classification logits $Y \in R^{B \times K}$

1. Initial Parametric Pooling:

Reduce node count by a factor of 64: $H \leftarrow \text{Linear}(X)$, where $H \in R^{B \times C \times N_{joints}}$ and $N_{joints} = N_{select} / 64$.

2. Dynamic Adjacency Matrix Construction:

Project features to queries and keys: $Q \leftarrow \text{Conv1d}(H)$, $K \leftarrow \text{Conv1d}(H)$.

Channel-wise mean pooling: $q, k \leftarrow \text{Mean}(Q, K, \text{dim} = \text{channel}) \in R^{B \times N}$.

Compute pairwise scalar affinity (via broadcasting): $A_{dyn} \leftarrow \tanh(q1^T - 1k^T)$.

Fuse with learnable prior A_{prior} : $A \leftarrow A_{prior} + \alpha \cdot A_{dyn}$ where α is a learnable scalar.

3. Graph Convolution & Message Passing:

Feature transformation: $H' \leftarrow \text{Conv1d}(H)$.

Graph aggregation: $H_{agg} \leftarrow H' \times A$.

Normalization: $H_{norm} \leftarrow \text{BatchNorm}(H_{agg})$.

4. Global Pooling & Classification:

Aggregate to global vector: $h_{global} \leftarrow \text{Linear}(H_{norm})$.

Regularization: $h_{global} \leftarrow \text{Dropout}(h_{global}, p = 0.1)$.

Final prediction: $Y \leftarrow \text{Linear}(h_{global})$.

Return Y

The classification loss, denoted as $\mathcal{L}_b$, is computed using the standard cross-entropy loss function. The total loss function $\mathcal{L}$ of the model, which combines the classification loss with the alignment loss, is formulated as follows:

$$\mathcal{L}_b = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^{C_k} y_{i,c} \log(p_{i,c}) \quad (13)$$

$$\mathcal{L} = \alpha \mathcal{L}_{\perp} + \mathcal{L}_b \quad (14)$$

where N denotes the total number of samples utilized for classification, $y_{i,c}$ represents the ground truth label for sample i in class c , and $p_{i,c}$ signifies the predicted probability that sample i belongs to class c .

3.2. Datasets

To evaluate the performance of the proposed model, we conducted experiments on two open-source datasets: tomato images from the PlantVillage dataset [38] (year 2015) and the 2018 AI Challenger dataset [39] (year 2018), from December 2025 to June 2026. After filtering for the tomato category, the PlantVillage dataset comprises 16,012 images spanning 9 disease classes and 1 healthy class. Similarly, the 2018 AI Challenger dataset contains 12,968 images across the 10 categories (9 diseases and 1 healthy). Both datasets were split into training, validation, and test sets following a 7:1:2 ratio. As detailed in Table 1 and Figure 5, both datasets exhibit varying degrees of class imbalance. Notably, while the PlantVillage dataset features relatively controlled backgrounds, the AI Challenger dataset presents highly complex, unstructured natural backgrounds. Consequently, the AI Challenger dataset serves as the primary benchmark for our in-depth ablation studies and

sensitivity analyses to rigorously validate the robustness of the proposed framework under challenging real-world conditions.

Table 1. Dataset Composition and Class Distribution.

Category	PlantVillage			2018 AI Challenger		
	Train	Val	Test	Train	Val	Test
Bacterial Spot	1488	213	426	-	-	-
Early Blight	699	101	200	555	79	158
Healthy	1113	160	318	967	138	276
Late Blight	1336	191	382	1074	153	307
Leaf Mold	666	96	190	528	76	151
Septoria Leaf Spot	1239	178	354	982	140	281
Two-spotted Spider Mite	1173	168	335	650	93	186
Target Spot	982	141	281	52	7	15
Tomato Mosaic Virus	261	38	74	208	30	60
Tomato Yellow Leaf Curl Virus	2246	321	642	3033	433	867
Powdery Mildew		-		1028	147	294
Total		16,012			12,968	

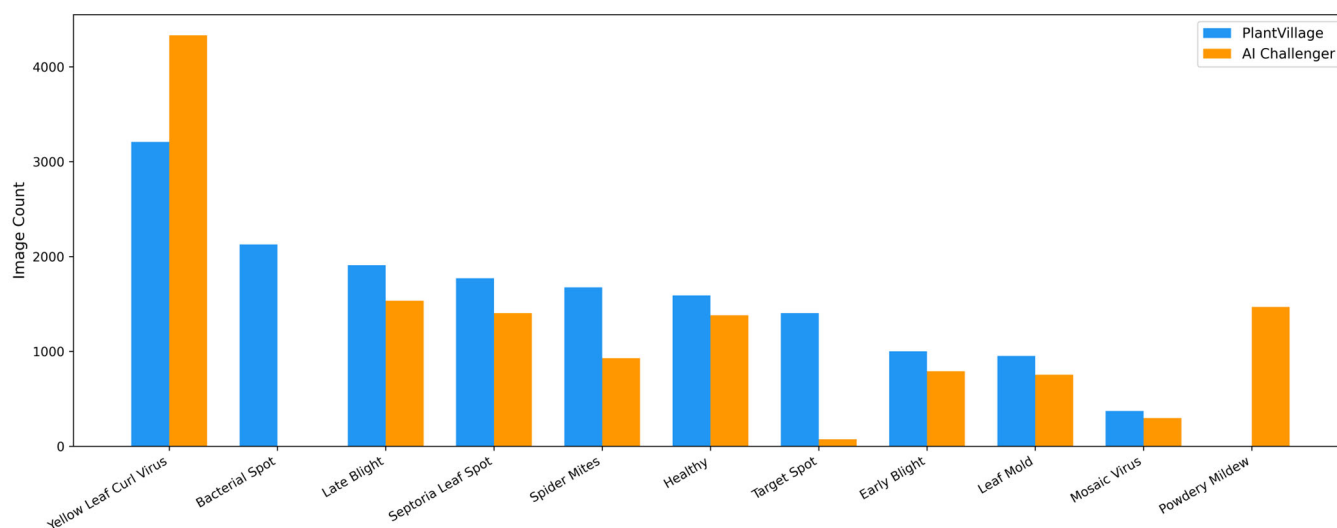


Figure 5. Distribution of image samples across different disease categories in the PlantVillage and 2018 AI Challenger datasets.

3.3. Experimental Setup

In our model, the parameter K is set to $\{256, 128, 64, 32\}$ to capture key points at different layers. The weighting factor α in the loss function is set to 0.4. The temperature parameter T , used for feature alignment, is annealed linearly from 10 to 1 during training.

The aforementioned hyperparameter settings are initially determined based on grid search on the validation set and task-specific priors regarding the spatial sparsity of tomato lesions. A comprehensive sensitivity analysis and granular ablation study regarding the impact of these parameters, as well as the individual contributions of other architectural components, are deeply investigated and discussed in Section 4.2.

Regarding data preprocessing, all input images undergo a standardized augmentation pipeline. Specifically, images are first resized such that the shorter side measures 256 pixels while preserving the original aspect ratio. Subsequently, a 224×224 region is randomly cropped. To enhance data diversity, horizontal flipping is applied with a probability of 0.5. Lastly, pixel arrays are transformed into tensor format and subjected to normalization based on the channel-wise statistics (mean and std) of the respective datasets.

All models were developed within the PyTorch 2.8.0 ecosystem and executed on a workstation equipped with an NVIDIA RTX 3080 Ti (12 GB) GPU. We employed the Adam optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-4} . A batch size of 128 was achieved via gradient accumulation. A Cosine Annealing scheduler was utilized to dynamically adjust the learning rate, with a minimum threshold of 1×10^{-5} . The models were trained for a total of 150 epochs.

To ensure a strictly fair and reproducible comparison, all classic baseline models (including VGG19, ResNet50, ResNet101, MobileNetV2, and DenseNet121) were re-implemented and trained under identical experimental conditions, utilizing the exact same data splits, augmentation pipelines, input resolutions, and optimization protocols. Furthermore, to guarantee statistical reliability and mitigate the variance introduced by random initialization, all main experiments and ablation studies were independently executed three times with different random seeds (42, 123, and 456). The final performance metrics are reported as the mean and standard deviation across these three runs.

3.4. Evaluation Metrics

Top-1 Accuracy is the most prevalent metric for classification tasks, defined as the proportion of samples where the predicted class with the highest probability matches the ground truth label. This metric provides an intuitive reflection of the model's overall performance in single-label classification scenarios.

$$\text{Top-1 Accuracy} = \frac{\text{Correct}}{\text{Total}} \quad (15)$$

Precision quantifies the accuracy of positive predictions. It represents the ratio of true positives to all samples predicted as positive. A higher precision indicates a lower rate of false positives (i.e., fewer non-diseased samples misclassified as diseased). Recall measures the model's coverage capability, defined as the proportion of actual positive samples that are correctly identified. A higher recall signifies a stronger ability to detect positive instances (i.e., fewer missed disease cases).

$$\text{Precision} = \frac{TP}{TP + FP} \quad (16)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (17)$$

Here, True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) are calculated individually for each class. Specifically, TP denotes the count of samples that belong to a specific class and are correctly predicted as such; TN represents samples neither belonging to nor predicted as that class; FP corresponds to instances falsely classified as members of the target category, whereas FN denotes ground-truth positives that are misclassified as negative.

Defined as the harmonic average of Precision and Recall, the F1-Score provides a unified measure of diagnostic accuracy. By reconciling the inherent conflict between these two indicators, it proves especially effective for handling skewed class distributions, which are prevalent in agricultural pathology. Notably, this score remains low unless both precision and recall achieve substantial values.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (18)$$

In this study, the F1-Score for each class is first computed based on its respective confusion matrix values (TP, TN, FP, FN). The final reported F1-Score is then derived as

the weighted average of individual class F1-Scores, where the weights correspond to the support (number of samples) of each class.

4. Results and Discussion

4.1. Comparison with State-of-the-Art Models

The primary evaluation of the proposed EA-HFA model was conducted on the PlantVillage tomato dataset.

To benchmark performance, we compared our approach against five classic deep learning architectures: VGG19 [18], ResNet50 [19], ResNet101 [19], MobileNetV2 [40], and DenseNet121 [20]. Furthermore, we extended the comparison to include existing state-of-the-art models specifically designed for tomato disease classification.

As shown in Tables 2 and 3, the proposed EA-HFA model achieves highly competitive performance, attaining a Top-1 accuracy of 99.29% and an F1-Score of 0.992. Notably, these results surpass those of most classical models and existing specialized models. The experimental results on the PlantVillage dataset demonstrate that our model exhibits exceptional robustness and discriminative capability in classifying tomato diseases characterized by extremely high inter-class visual similarity.

Table 2. Performance Comparison with Classic Models on the PlantVillage Dataset.

Model	Top-1 (%)	Macro F1	Weighted F1	Macro Precision	Macro Recall	Model Size (MB)
VGG19	99.85 ± 0.15	0.987 ± 0.002	0.999 ± 0.002	0.987 ± 0.001	0.987 ± 0.001	532
MobileNetV2	98.61 ± 0.11	0.985 ± 0.002	0.986 ± 0.001	0.985 ± 0.001	0.985 ± 0.001	9
DenseNet121	99.17 ± 0.23	0.991 ± 0.002	0.991 ± 0.002	0.991 ± 0.002	0.991 ± 0.002	27
ResNet50	99.43 ± 0.07	0.993 ± 0.001	0.994 ± 0.001	0.993 ± 0.001	0.994 ± 0.001	90
ResNet101	99.45 ± 0.14	0.994 ± 0.002	0.994 ± 0.002	0.994 ± 0.001	0.994 ± 0.001	162
Ours	99.29 ± 0.19	0.992 ± 0.002	0.987 ± 0.002	0.993 ± 0.002	0.992 ± 0.002	124

Table 3. Performance Comparison with SOTA Models on the PlantVillage Dataset.

Model	Top-1 (%)	Model Size (MB)
Tm et al. [29]	94.85	156.78
Ahmed et al. [41]	99.30	9.6
Agarwal et al. [42]	98.40	0.208
Abbas et al. [43]	97.11	27.58
Zhichao Chen et al. [29]	99.45	29
Ours	99.29 ± 0.19	124

However, it is worth noting that our EA-HFA model yields a slightly lower Top-1 accuracy on the PlantVillage dataset compared to ResNet50 and ResNet101. This marginal performance gap can be primarily attributed to the inherent characteristics of the PlantVillage dataset, which features overly simple and homogeneous laboratory-controlled backgrounds. The core assumption of our Entropy Attention mechanism is that sparse lesion regions yield the lowest entropy. However, in highly controlled environments with uniform backgrounds, the background pixels may also exhibit extremely low entropy, occasionally misleading the attention mechanism or causing the Top-K hard selection to over-filter useful global context. In contrast, traditional architectures like ResNet50 and ResNet101, which rely on global pooling, can effortlessly capture holistic features without the risk of information loss in such simplified scenarios. Nevertheless, the primary design motivation of our proposed framework is to address the spatial sparsity of discriminative features and severe background interference in complex, unstructured real-world environments.

Therefore, while incurring a minor accuracy trade-off in overly simplified scenarios, its practical utility and effectiveness in filtering background noise are further evaluated on the significantly more challenging AI Challenger dataset in the following.

Table 4 presents the Macro Precision, Recall, and F1-Score for each class on the PlantVillage dataset. Notably, the Precision and F1-Score for the vast majority of categories exceed 0.99. This consistent high performance across all classes underscores the model's robustness and reliability, particularly in class-imbalanced scenarios. It demonstrates that the proposed method maintains exceptional accuracy and generalization capability, effectively mitigating the bias towards majority classes often observed in imbalanced datasets.

Table 4. Per-Class Performance Metrics on the PlantVillage Dataset.

Class	Macro Precision (%)	Macro Recall (%)	Macro F1-Score (%)
Bacterial Spot	99.07 $\pm$ 0.23	99.61 $\pm$ 0.14	99.34 $\pm$ 0.07
Early Blight	98.99 $\pm$ 0.85	97.50 $\pm$ 1.73	98.23 $\pm$ 0.77
Healthy	99.90 $\pm$ 0.18	99.16 $\pm$ 0.48	99.53 $\pm$ 0.16
Late Blight	99.14 $\pm$ 0.65	99.74 $\pm$ 0.26	99.44 $\pm$ 0.40
Leaf Mold	99.31 $\pm$ 0.79	99.82 $\pm$ 0.30	99.56 $\pm$ 0.30
Septoria Leaf Spot	99.53 $\pm$ 0.58	99.34 $\pm$ 0.71	99.43 $\pm$ 0.37
Two-spotted Spider Mite	98.14 $\pm$ 0.66	99.40 $\pm$ 0.30	98.76 $\pm$ 0.22
Target Spot	99.52 $\pm$ 0.54	98.34 $\pm$ 0.41	98.93 $\pm$ 0.31
Tomato Mosaic Virus	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00
Tomato Yellow Leaf Curl Virus	99.64 $\pm$ 0.09	99.53 $\pm$ 0.16	99.58 $\pm$ 0.09

Figure 6 depicts the classification performance via a confusion matrix for the PlantVillage benchmark. It is evident that the vast majority of samples are correctly classified along the diagonal corresponding to their ground truth labels, with negligible misclassification errors. This near-perfect alignment further substantiates the model's robust generalization capability, demonstrating its effectiveness in handling multi-class plant disease recognition tasks across varying data distributions.

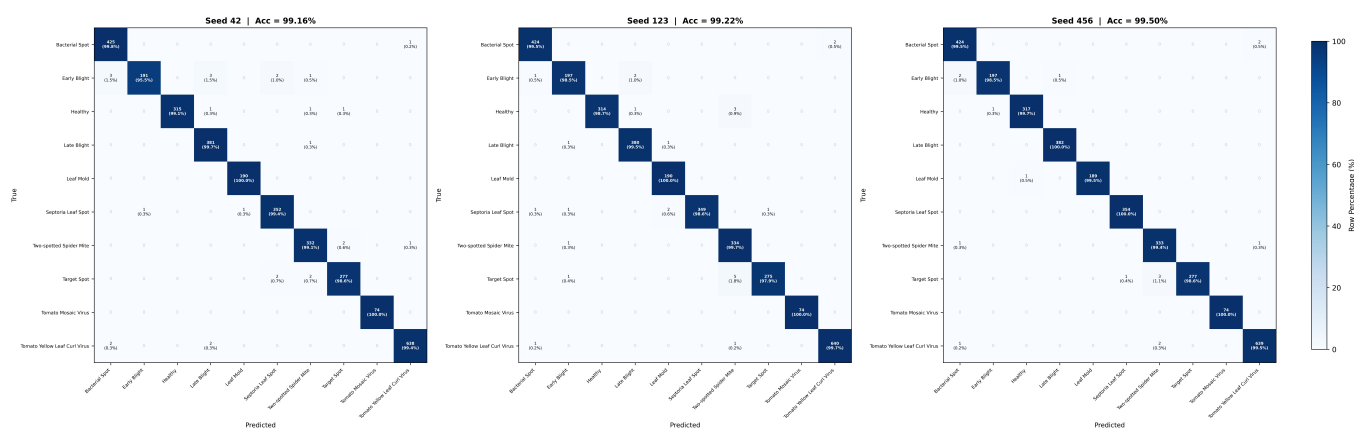


Figure 6. Normalized confusion matrix of the proposed EA-HFA model's classification performance on the PlantVillage dataset.

To fully evaluate the performance of our method, we conducted the main experiments on the 2018 AI Challenger tomato dataset. Compared to the controlled environment of PlantVillage, this dataset features more complex and realistic backgrounds, enabling a more accurate reflection of our method's performance. We benchmarked the proposed EA-HFA against five classic deep learning architectures: VGG19 [18], ResNet50 [19], ResNet101 [19], MobileNetV2 [40] and DenseNet121 [20].

As shown in Table 5, EA-HFA achieves a Top-1 accuracy of $97.82\% \pm 0.29$, yielding performance statistically comparable to established deep learning architectures such as ResNet50 ($97.96\% \pm 0.41$) and ResNet101 ($97.96\% \pm 0.29$), with the mean difference (0.14%) falling well within the margin of standard deviation. This result indicates that, under complex natural backgrounds, our framework maintains discriminative capability on par with mainstream CNNs while pursuing a more interpretable feature selection mechanism. Notably, our method maintains a Macro Precision of 0.961 ± 0.007 and a Macro F1-Score of 0.959 ± 0.002 , suggesting balanced performance across categories with varying sample sizes. These metrics, together with the relatively lower standard deviation observed in our model, imply a certain degree of training stability under stochastic initialization, highlighting its practical utility for future field-oriented applications in unstructured agricultural scenarios.

Table 5. Performance Comparison on the 2018 AI Challenger Dataset.

Model	Top-1 (%)	Macro F1	Weighted F1	Macro Precision	Macro Recall	Model Size (MB)
VGG19	97.51 ± 0.26	0.957 ± 0.001	0.975 ± 0.003	0.960 ± 0.010	0.955 ± 0.008	532
MobileNetV2	97.23 ± 0.04	0.941 ± 0.005	0.972 ± 0.001	0.943 ± 0.011	0.941 ± 0.002	9
DenseNet121	97.80 ± 0.15	0.956 ± 0.010	0.978 ± 0.001	0.958 ± 0.006	0.955 ± 0.014	27
ResNet50	97.96 ± 0.41	0.956 ± 0.007	0.979 ± 0.004	0.958 ± 0.009	0.956 ± 0.003	90
ResNet101	97.96 ± 0.33	0.961 ± 0.007	0.979 ± 0.003	0.964 ± 0.006	0.958 ± 0.008	162
Ours	97.82 ± 0.29	0.959 ± 0.002	0.978 ± 0.003	0.961 ± 0.007	0.957 ± 0.003	124

Table 6 details the per-class Macro Precision, Recall, and F1 metrics evaluated on the 2018 AI Challenger dataset. Notably, the F1-Score for the majority of categories remains above 0.95, contributing to a balanced Macro F1-Score of 0.959. However, for the most challenging small-sample category, Target Spot (with only 15 test images), the model yields a relatively lower F1-Score of 0.846 (similarly, Early Blight achieves 0.905). This performance drop transparently reflects the inherent difficulty of learning robust feature representations for extremely rare classes under severe class imbalance and complex background interference. Nevertheless, the overall per-class metrics indicate that the proposed framework maintains reliable discriminative capability across most categories, providing a practical baseline for real-world agricultural monitoring where data distributions are inherently skewed.

Table 6. Per-Class Performance Metrics on the 2018 AI Challenger Dataset.

Class	Macro Precision (%)	Macro Recall (%)	Macro F1-Score (%)
Early Blight	93.38 ± 1.91	87.84 ± 4.76	90.46 ± 2.26
Healthy	99.39 ± 0.41	98.31 ± 1.11	98.84 ± 0.43
Late Blight	94.95 ± 1.91	97.50 ± 0.68	96.20 ± 1.18
Leaf Mold	98.21 ± 0.37	96.91 ± 1.01	97.55 ± 0.40
Septoria Leaf Spot	97.51 ± 1.61	97.15 ± 0.71	97.33 ± 1.07
Two-spotted Spider Mite	95.75 ± 2.39	95.88 ± 1.73	95.80 ± 1.46
Target Spot	85.01 ± 8.22	84.44 ± 3.85	84.55 ± 4.48
Tomato Mosaic Virus	97.86 ± 2.42	99.44 ± 0.96	98.63 ± 1.25
Tomato Yellow Leaf Curl Virus	99.20 ± 0.30	99.81 ± 0.18	99.50 ± 0.14
Powdery Mildew	99.77 ± 0.20	99.89 ± 0.20	99.83 ± 0.17

Figure 7 presents the corresponding confusion matrix. Consistent with the observations on PlantVillage, the vast majority of samples are correctly classified along the diagonal representing ground truth labels, with only a negligible number of misclassifications.

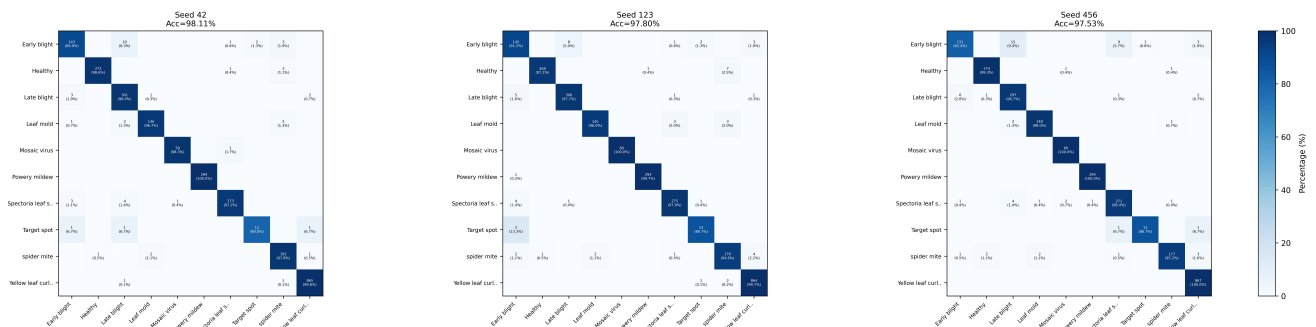


Figure 7. Normalized confusion matrix illustrating the classification results of the EA-HFA model on the 2018 AI Challenger dataset. In the figure, “Septoria Leaf S...” stands for Septoria Leaf Spot, and “Yellow Leaf Cur...” stands for Yellow Leaf Curl Virus.

To further evaluate the training dynamics and stability of the proposed EA-HFA framework, Figure 8 illustrates the training loss curves on both the PlantVillage and 2018 AI Challenger datasets across three independent random seeds (42, 123, and 456). As observed, the loss values exhibit a rapid and steady descent during the initial epochs, followed by a smooth convergence to a stable minimum. The training trajectories for the three distinct seeds show a high degree of overlap with negligible variance on both datasets. This remarkable consistency not only corroborates the robustness of our architecture against random initialization but also guarantees the statistical reliability and reproducibility of the reported performance metrics, indicating that the model’s consistent performance stems from its intrinsic structural design rather than stochastic variance.

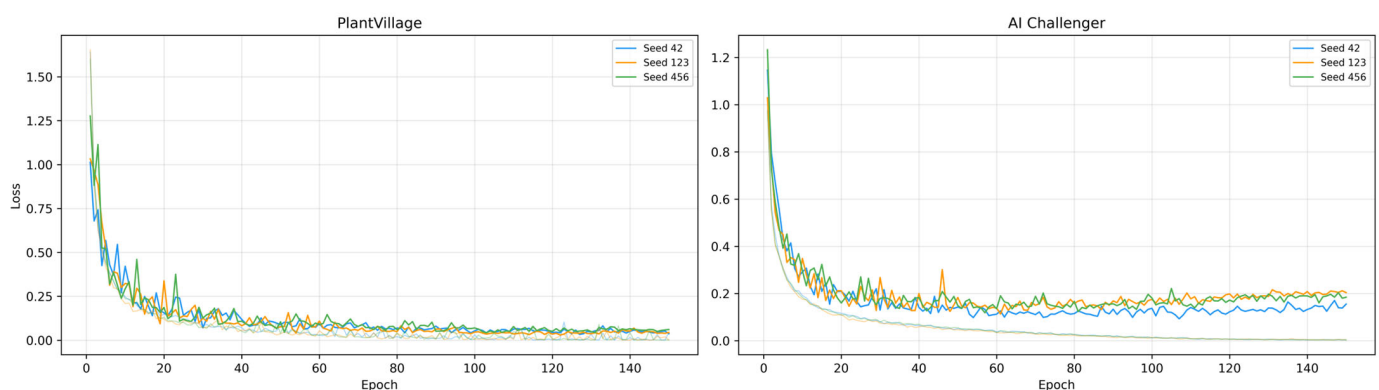


Figure 8. Training loss curves of the EA-HFA model on two datasets. The solid line represents the validation set loss, and the faded line represents the training set loss.

Collectively, the experimental results across both the PlantVillage and 2018 AI Challenger datasets confirm the model’s ability to maintain reliable discriminative capability and low misclassification rates across diverse plant disease data with distinct distributions. This stability underscores the model’s practical utility across varying scenarios and highlights its significant potential as a reliable alternative for future field-oriented agricultural applications.

To evaluate the computational efficiency and practical deployment potential of the proposed framework, we conducted a comprehensive complexity analysis. The benchmarking equipment, environment, and data preprocessing methods followed the same experimental setup as in previous experiments. Latency measurement included only model forward propagation, excluding image decoding, data loading, and preprocessing transforms, with 50 warm-up iterations and 300 timed repeats. We also measured the parameter count

(Params), calculated floating-point operations (FLOPs) using the thop library, and recorded peak GPU memory allocation.

As summarized in Table 7, our EA-HFA model requires 32.59 M parameters and 6.88 G FLOPs. Notably, compared to the heavier ResNet101, our model achieves approximately a 23% reduction in parameters and around a 12% reduction in FLOPs, while maintaining comparable inference latency and lower peak memory usage. This remarkable efficiency is primarily attributed to the proposed Entropy Attention mechanism. By explicitly discarding non-discriminative background regions via hard selection, it reduces the computational burden on subsequent deeper layers and the Graph Convolutional Network (GCN) classifier. Although lightweight models like MobileNetV2 exhibit the lowest computational cost, their accuracy on the challenging AI Challenger dataset is noticeably lower. Overall, EA-HFA strikes a reasonable balance between high-precision fine-grained recognition and computational efficiency.

Table 7. Computational Efficiency Comparison of Different Models.

Model	Params (M)	FLOPs (G)	Latency bs = 1 (ms/img)	Peak Allocated bs = 1 (MB)
VGG19	139.611	19.628	3.177	566.2
DenseNet121	6.964	2.896	16.642	46.1
ResNet101	42.521	7.864	12.242	183.2
MobileNetV2	2.237	0.326	5.342	27.5
Ours	32.588	6.88	12.103	161.8

4.2. Ablation Studies

To comprehensively evaluate the contribution of each proposed component, we conducted a granular ablation study on the challenging AI Challenger dataset, as detailed in Table 8. In the table, “Feature Fusion” explicitly refers to the PANet-based feature aggregation module. Specifically, when only the Feature Fusion module is introduced to the Backbone, the model’s accuracy drops noticeably from 97.96% to 97.51%. Furthermore, adding the Hierarchical Feature Alignment (HFA) without Entropy Attention yields a statistically comparable but still degraded accuracy of $97.51\% \pm 0.34$. This counter-intuitive phenomenon reveals a critical insight: in complex, natural backgrounds, blindly fusing multi-scale features or enforcing cross-layer alignment without explicit target guidance forces the network to inadvertently align background noise and irrelevant textures.

Table 8. Ablation Study of Model Components.

Backbone	Feature Fusion (PANet)	Hierarchical Feature Alignment	Entropy Attention	TOP-1 (%)
✓	×	×	×	97.96 ± 0.41
✓	✓	×	×	97.51 ± 0.19
✓	✓	✓	×	97.51 ± 0.34
✓	✓	×	✓	97.86 ± 0.37
✓	✓	✓	✓	97.82 ± 0.29

Symbol ‘✓’ indicates module is included; ‘×’ indicates it is excluded.

However, the introduction of the EA module (Feature Fusion + Entropy) dramatically recovers the performance to $97.86\% \pm 0.37$ by explicitly filtering out background noise and focusing on sparse discriminative lesions, which verifies the core effectiveness of our information-theoretic hard selection mechanism. Finally, the Full Model achieves a Top-1 accuracy of $97.82\% \pm 0.29$. Although this mean accuracy is marginally lower than that of

the EA-only variant, the difference falls well within the margin of standard deviation. This indicates that the HFA module, while not yielding a further marginal gain in global Top-1 accuracy, functions as a supplementary constraint to encourage cross-scale consistency, thereby helping to maintain training stability (reflected in a lower standard deviation) and balanced category performance.

Table 9 further isolates the impact of structural designs and hyperparameters. Interestingly, reducing the selection capacity K by half $\{128,64,32,16\}$ yields a marginally higher mean accuracy ($97.88\% \pm 0.21$) compared to the full model ($97.82\% \pm 0.29$). This result aligns with our theoretical assumption that lower-entropy pixels contain more discriminative information, demonstrating that the entropy-based hard selection can effectively filter out background noise. However, overly aggressive selection may risk discarding subtle spatial contexts. Therefore, the optimal determination of K for varying disease scales remains a subject for our future investigation.

Table 9. Impact of Different Strategies on Model Performance.

Setting	TOP-1 (%)
<i>w/o</i> Position Embedding	97.71 ± 0.35
Reduced K	97.88 ± 0.21
<i>w/o</i> GCN	97.57 ± 0.07
$T = 1$	97.80 ± 0.34
$T = 10$	97.73 ± 0.13
Ours	97.82 ± 0.29

Regarding positional embeddings, removing them results in a noticeable performance decrease (dropping to $97.71\% \pm 0.35$). This confirms their theoretical necessity for constructing the structured spatial relationships required by the subsequent GCN classifier. Without explicit spatial priors, the GCN struggles to effectively model the intrinsic topological dependencies among the sparse discriminative nodes, highlighting the effectiveness of positional embeddings in our current architecture. Regarding the temperature T , fixing it at constant extremes ($T = 1$ or $T = 10$) yields statistically comparable mean accuracies to our annealing strategy. However, the dynamic annealing is theoretically grounded to prevent early-stage overconfidence and ensure a smooth transition from learning soft inter-class relationships to focusing on hard discriminative boundaries, yielding more reliable convergence dynamics.

As detailed in Table 9, notably, to isolate the specific impact of the GCN module, we replaced it with a standard Fully Connected (FC) layer (denoted as “*w/o* GCN”). The results show that removing the GCN leads to a performance degradation (the average accuracy drops to $97.57\% \pm 0.07$, compared to $97.82\% \pm 0.29$ of the full model). This degradation occurs because standard FC layers treat the aggregated multi-scale features as a flat, unordered vector, thereby discarding the intrinsic topological relationships and spatial affinities among the sparse discriminative nodes. In contrast, the GCN effectively preserves and reasons over these structural dependencies through dynamic message passing, proving its indispensability in handling complex spatial structures.

Figure 9 visualizes the entropy maps across different hierarchical layers after training the model with low-entropy and mean-entropy strategies, respectively. These results demonstrate the capability of the proposed low-entropy attention mechanism to help identify discriminative features across multiple scales.

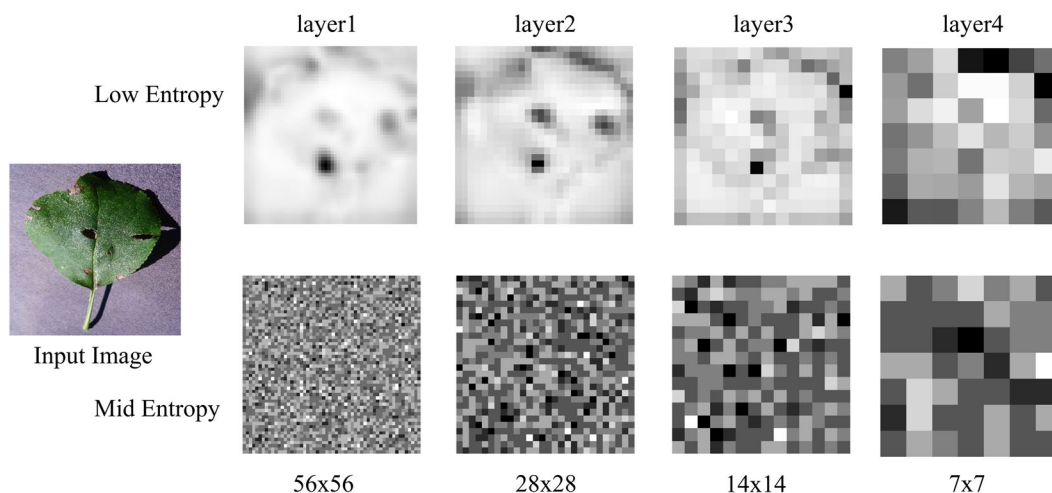


Figure 9. Visualization of entropy maps generated at different hierarchical layers of the model. Darker regions correspond to lower entropy values. The ‘Low Entropy’ row demonstrates that the proposed attention mechanism tends to identify and focus on discriminative lesion areas across multiple scales, from layer 1 (56×56) to layer 4 (7×7).

4.3. Visualization

Figure 10 visualizes the attention maps generated by the Entropy Attention module at each layer and compares them with the Grad-CAM results of various baseline models across different blocks. These visualizations intuitively reveal the specific image regions focused on by different layers of each model.

As observed in the sub-figures for ResNet50, most baseline models tend to focus on low-level details such as textures in their shallow layers. However, these local features often fail to concentrate on the discriminative regions or features critical for image classification [44]. This observation corroborates the limitations of traditional image classification models when dealing with tasks characterized by sparse feature distributions. Furthermore, in deeper layers, traditional models either drift away from the discriminative regions entirely or attend to areas that are insufficiently comprehensive, as seen in the ResNet101 sub-figures.

In contrast, benefiting from the Entropy Attention mechanism, our model tends to concentrate on these critical features even in shallow layers. For instance, as shown in the EA-HFA sub-figure, the model focuses predominantly on the upper-middle section of the leaf, helping to mitigate interference from background noise. In the deeper layers of our model, while the attention coverage becomes more extensive to capture global context, the Hierarchical Feature Alignment module helps maintain the focus on regions. This design aims to compensate for potential omissions in shallow-layer focusing while providing rich global contextual information. Ultimately, our model’s architecture automatically synthesizes features from all four hierarchical layers to make the final classification decision. Rather than claiming a distinct superiority in overall accuracy, this multi-scale feature synthesis offers an interpretable alternative perspective, illustrating how the model explicitly prioritizes lesion-relevant areas throughout its decision-making process.

Figure 11 illustrates the distribution of features from the final layer of our model after dimensionality reduction using t-SNE. In the learned feature space, the vast majority of classes form distinct and compact clusters with clear intra-class cohesion. Notably, even for confusable categories that exhibit high similarity in macroscopic morphology, the model achieves well-defined decision boundaries.

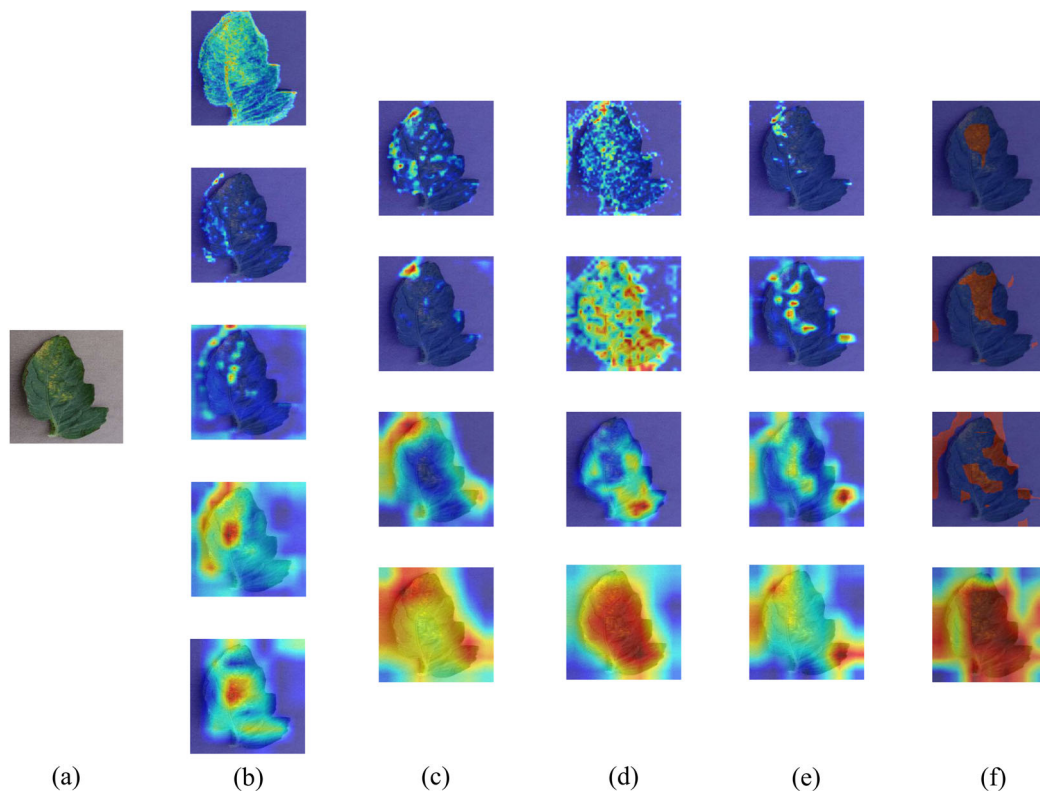


Figure 10. Comparative visualization of attention maps. Column (a) is the input image. Columns (b–e) show Grad-CAM heatmaps for baseline models (VGG19, DenseNet121, ResNet50, ResNet101). Column (f) displays the attention map from our proposed EA-HFA model. To intuitively highlight the model's focus regions, the heatmaps were generated by applying a value threshold based on the proportion of selected feature points relative to the total feature map size at each layer. The comparison reveals that baseline models often focus on irrelevant background or textures, while our model (f) appears to better concentrate on lesion-relevant areas, demonstrating superior focus on discriminative regions. The heatmap overlays the original image, where warmer colors (red) indicate higher activation, and cooler colors (blue) indicate lower activation.

These observations suggest that the proposed model is capable of realizing inter-class discriminability within the complex domain of tomato leaf pathology identification. This structural clarity in the feature space aligns with the aforementioned quantitative metrics, indicating that the model has learned highly discriminative feature representations. Specifically, it suggests that the Entropy Attention mechanism helps filter out background noise and non-discriminative texture interference, enabling the model to focus more consistently on the genuine discriminative features of the disease lesions, thereby constructing a more separable feature representation.

4.4. Discussion

While the proposed EA-HFA framework demonstrates strong performance and robust feature representation, as evidenced by the quantitative and qualitative results above, several critical aspects regarding its applicable conditions, limitations, and practical deployment warrant in-depth discussion.

First, regarding the structural design, our ablation study confirms the necessity of absolute positional embeddings for GCN-based spatial modeling. However, in highly unstructured wild environments, absolute spatial priors may still be susceptible to complex background interference. This suggests that exploring adaptive or relative positional encoding mechanisms (e.g., Rotary Position Embedding) could be a valuable direction for future research to further enhance robustness in real-field scenarios.

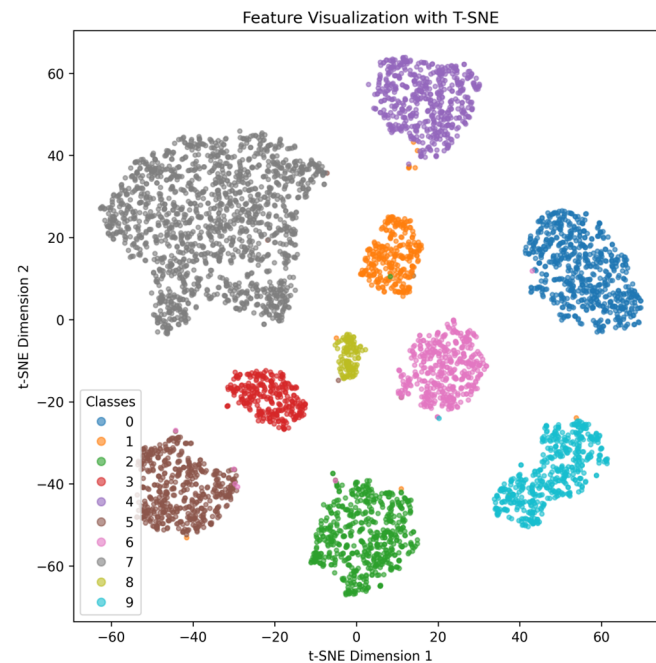


Figure 11. Feature Visualization using t-SNE. Each colored cluster represents a distinct tomato disease class.

Second, regarding the low-entropy attention mechanism, its core assumption is that lesion regions yield the highest predictive confidence (i.e., the lowest entropy). However, when lesions become excessively large (covering most of the leaf) or the background is overly simple and homogeneous, failure cases may arise. In such extreme scenarios, the entropy distribution may become uniform or biased, causing the mechanism to mistakenly attend to non-discriminative regions.

Third, the model exhibits noticeable limitations on small-sample categories. For instance, in the AI Challenger dataset, the recognition performance on the Target Spot class, which has an extremely limited sample size (only 52 training and 15 test images), is relatively lower (achieving an F1-Score of 0.846). The severe class imbalance and the lack of diverse morphological variations hinder the model's ability to learn robust feature representations for these rare diseases, indicating that the current framework still relies heavily on sufficient data support and underscoring the necessity of incorporating few-shot learning or advanced data augmentation techniques in future work.

Fourth, a significant domain gap exists between public datasets and real-field environments. Datasets such as PlantVillage and AI Challenger primarily feature images captured under controlled or semi-controlled conditions. In contrast, real-world agricultural deployment faces unpredictable challenges such as lighting changes and leaf occlusion [45], highly complex backgrounds [46], and varying camera angles, which potentially amendable by exposure-aware super-resolution, contrastive learning, and multi-scale convolution [47–50]. The current reliance on the ResNet50 backbone further constrains mobile deployment.

Finally, regarding practical deployment in smart agriculture, while our efficiency analysis demonstrates that EA-HFA maintains a favorable computational footprint on standard GPUs, we acknowledge a limitation regarding edge-device validation. Real-world agricultural monitoring often relies on low-resource edge devices (e.g., UAVs or handheld sensors) with strict power and latency constraints. Since our current benchmarking was exclusively conducted on a high-performance desktop GPU, the actual inference speed and memory footprint on edge hardware may vary due to different memory access costs and hardware architectures. Therefore, adapting the framework to lightweight backbones

(e.g., MobileNetV3), applying model quantization or pruning, and conducting rigorous on-device testing in actual field conditions remain critical directions for our future work.

5. Conclusions

To address the core challenges of fine-grained tomato leaf disease classification—namely, the spatial sparsity of discriminative features, severe background noise interference, and multi-scale feature inconsistency—we introduce EA-HFA, an innovative framework that couples an EA strategy and a HFA component in a unified manner.

The primary contributions of this work are twofold: (1) We introduce an information-theoretic Entropy Attention mechanism that quantifies predictive uncertainty to adaptively focus on sparse yet critical lesion regions. This helps mitigate irrelevant background interference and alleviates the signal dilution problem prevalent in conventional models. (2) We design a Hierarchical Feature Alignment module that leverages knowledge distillation and KL-divergence constraints to encourage semantic consistency across network depths, thereby improving the model's multi-scale perceptual stability for subtle pathological features.

Comprehensive experimental evaluations demonstrate that EA-HFA achieves superior performance on both the PlantVillage and AI Challenger 2018 datasets, attaining Top-1 accuracies of 99.29% and 97.82%, respectively, and achieving results statistically comparable to mainstream deep learning architectures while offering enhanced interpretability. Extensive ablation studies and qualitative analyses further corroborate the individual efficacy and synergistic interaction of the proposed modules, validating the practical utility and architectural rationality of EA-HFA.

Future work will address these challenges by (1) exploring adaptive entropy thresholds to mitigate failure cases under extreme lesion scales; (2) introducing few-shot learning or advanced data augmentation strategies to alleviate the small-sample bottleneck; and (3) adapting the framework to lightweight architectures (e.g., MobileNetV3) and validating it on large-scale, in-the-wild field imagery to bridge the domain gap and support practical smart agriculture systems.

Author Contributions: Conceptualization, Z.S.; methodology, Z.S. and J.W.; formal analysis, J.W.; validation, S.Y. and B.F.; Writing—original draft, Z.S. and S.Y.; Writing—review and editing, J.W. and B.F.; supervision, S.Y. and B.F.; funding acquisition, Z.S. and S.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Undergraduate Innovation and Entrepreneurship Training Program of China (No. 202511057064) and the “Pioneer” and “Leading Goose” R&D Program of Zhejiang (No. 2024C01109). And we thank Jiantao Sun and Mengling Mao for their help.

Data Availability Statement: The data presented in this study are openly available in Kaggle at <https://www.kaggle.com/datasets/abdallahalidev/plantvillage-dataset> (accessed on 23 June 2026), reference number [38], and in Aistudio at <https://aistudio.baidu.com/datasetdetail/199888> (accessed on 23 June 2026), reference number [39]. The code in this study is openly available in GitHub at <https://github.com/Mirage2026/EA-FHA> (accessed on 23 June 2026).

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

References

1. Panno, S.; Davino, S.; Caruso, A.G.; Bertacca, S.; Crnogorac, A.; Mandić, A.; Noris, E.; Matić, S. A Review of the Most Common and Economically Important Diseases That Undermine the Cultivation of Tomato Crop in the Mediterranean Basin. *Agronomy* **2021**, *11*, 2188. [CrossRef]
2. Xie, C.; Shao, Y.; Li, X.; He, Y. Detection of Early Blight and Late Blight Diseases on Tomato Leaves Using Hyperspectral Imaging. *Sci. Rep.* **2015**, *5*, 16564. [CrossRef] [PubMed]

3. Tang, Z.; He, X.; Zhou, G.; Chen, A.; Wang, Y.; Li, L.; Hu, Y. A Precise Image-Based Tomato Leaf Disease Detection Approach Using PLPNet. *Plant Phenomics* **2023**, *5*, 0042. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Zhao, J.; Xu, L.; Ma, Z.; Li, J.; Wang, X.; Liu, Y.; Du, X. A Review of Plant Leaf Disease Identification by Deep Learning Algorithms. *Front. Plant Sci.* **2025**, *16*, 1637241. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Yang, S.; Yang, X.; Wu, J.; Feng, B. Significant feature suppression and cross-feature fusion networks for fine-grained visual classification. *Sci. Rep.* **2024**, *14*, 24051. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Wei, X.-S.; Song, Y.-Z.; Aodha, O.M.; Wu, J.; Peng, Y.; Tang, J.; Yang, J.; Belongie, S. Fine-Grained Image Analysis With Deep Learning: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 8927–8948. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Yang, T.; Wei, J.; Xiao, Y.; Wang, S.; Tan, J.; Niu, Y.; Duan, X.; Pan, F.; Pu, H. LT-DeepLab: An Improved DeepLabV3+ Cross-Scale Segmentation Algorithm for Zanthoxylum Bungeanum Maxim Leaf-Trunk Diseases in Real-World Environments. *Front. Plant Sci.* **2024**, *15*, 1423238. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Zhang, B.; Zhao, Q.; Feng, W.; Lyu, S. AlphaMEX: A Smarter Global Pooling Method for Convolutional Neural Networks. *Neurocomputing* **2018**, *321*, 36–48. [\[CrossRef\]](#)
9. Olshausen, B.A.; Field, D.J. Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images. *Nature* **1996**, *381*, 607–609. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Olshausen, B.A.; Field, D.J. Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1? *Vis. Res.* **1997**, *37*, 3311–3325. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Santana, A.; Colombini, E. Neural Attention Models in Deep Learning: Survey and Taxonomy. *arXiv* **2021**, arXiv:2112.05909.
12. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [\[CrossRef\]](#)
13. Al Bashish, D.; Braik, M.; Bani-Ahmad, S. Detection and Classification of Leaf Diseases Using K-Means-Based Segmentation and Neural-Networks-Based Classification. *Inf. Technol. J.* **2011**, *10*, 267–275. [\[CrossRef\]](#)
14. Al Hiary, H.; Bani Ahmad, S.; Reyalat, M.; Braik, M.; ALRahamneh, Z. Fast and Accurate Detection and Classification of Plant Diseases. *Int. J. Comput. Appl.* **2011**, *17*, 31–38. [\[CrossRef\]](#)
15. Tian, Y.; Zhao, C.; Lu, S.; Guo, X. Multiple Classifier Combination for Recognition of Wheat Leaf Diseases. *Intell. Autom. Soft Comput.* **2011**, *17*, 519–529. [\[CrossRef\]](#)
16. Li, D.; Yin, Z.; Zhao, Y.; Zhao, W.; Li, J. MLFAnet: A Tomato Disease Classification Method Focusing on OOD Generalization. *Agriculture* **2023**, *13*, 1140. [\[CrossRef\]](#)
17. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [\[CrossRef\]](#)
18. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
20. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
21. Borhani, Y.; Khoramdel, J.; Najafi, E. A Deep Learning Based Approach for Automated Plant Disease Classification Using Vision Transformer. *Sci. Rep.* **2022**, *12*, 11554. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Brahimi, M.; Boukhalfa, K.; Moussaoui, A. Deep Learning for Tomato Diseases: Classification and Symptoms Visualization. *Appl. Artif. Intell.* **2017**, *31*, 299–315. [\[CrossRef\]](#)
23. Sardogan, M.; Tuncer, A.; Ozen, Y. Plant Leaf Disease Detection and Classification Based on CNN with LVQ Algorithm. In Proceedings of the 2018 3rd International Conference on Computer Science and Engineering (UBMK), Sarajevo, Bosnia, 20–23 September 2018; pp. 382–385.
24. Rangarajan, A.K.; Purushothaman, R.; Ramesh, A. Tomato Crop Disease Classification Using Pre-Trained Deep Learning Algorithm. *Procedia Comput. Sci.* **2018**, *133*, 1040–1047. [\[CrossRef\]](#)
25. Maeda-Gutiérrez, V.; Galván-Tejada, C.E.; Zanella-Calzada, L.A.; Celaya-Padilla, J.M.; Galván-Tejada, J.I.; Gamboa-Rosales, H.; Luna-García, H.; Magallanes-Quintanar, R.; Guerrero Méndez, C.A.; Olvera-Olvera, C.A. Comparison of Convolutional Neural Network Architectures for Classification of Tomato Plant Diseases. *Appl. Sci.* **2020**, *10*, 1245. [\[CrossRef\]](#)
26. Sanida, M.V.; Sanida, T.; Sideris, A.; Dasygenis, M. An Efficient Hybrid CNN Classification Model for Tomato Crop Disease. *Technologies* **2023**, *11*, 10. [\[CrossRef\]](#)
27. Attallah, O. Tomato Leaf Disease Classification via Compact Convolutional Neural Networks with Transfer Learning and Feature Selection. *Horticulturae* **2023**, *9*, 149. [\[CrossRef\]](#)
28. Sun, Y.; Ning, L.; Zhao, B.; Yan, J. Tomato Leaf Disease Classification by Combining EfficientNetv2 and a Swin Transformer. *Appl. Sci.* **2024**, *14*, 7472. [\[CrossRef\]](#)
29. Chen, Z.; Wang, G.; Lv, T.; Zhang, X. Using a Hybrid Convolutional Neural Network with a Transformer Model for Tomato Leaf Disease Detection. *Agronomy* **2024**, *14*, 673. [\[CrossRef\]](#)

30. Ji, R.; Wen, L.; Zhang, L.; Du, D.; Wu, Y.; Zhao, C.; Liu, X.; Huang, F. Attention Convolutional Binary Neural Tree for Fine-Grained Visual Categorization. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 10465–10474.
31. Lu, J.; Zhang, W.; Zhao, Y.; Sun, C. Image Local Structure Information Learning for Fine-Grained Visual Classification. *Sci. Rep.* **2022**, *12*, 19205. [[CrossRef](#)] [[PubMed](#)]
32. Chou, P.-Y.; Kao, Y.-Y.; Lin, C.-H. Fine-Grained Visual Classification with High-Temperature Refinement and Background Suppression. *arXiv* **2023**, arXiv:2303.06442.
33. Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *arXiv* **2015**, arXiv:1503.02531.
34. Zagoruyko, S.; Komodakis, N. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. *arXiv* **2016**, arXiv:1612.03928.
35. Zhang, L.; Bao, C.; Ma, K. Self-Distillation: Towards Efficient and Compact Neural Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 4388–4403. [[CrossRef](#)] [[PubMed](#)]
36. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
37. Bera, A.; Wharton, Z.; Liu, Y.; Bessis, N.; Behera, A. SR-GNN: Spatial Relation-Aware Graph Neural Network for Fine-Grained Image Categorization. *IEEE Trans. Image Process.* **2022**, *31*, 6017–6031. [[CrossRef](#)] [[PubMed](#)]
38. Hughes, D.P.; Salathe, M. An Open Access Repository of Images on Plant Health to Enable the Development of Mobile Disease Diagnostics. *arXiv* **2015**, arXiv:1511.08060.
39. Nagabhushan, G. Plant Disease Recognition AI Challenger (PDR2018). 2026. Available online: <https://www.kaggle.com/datasets/nagabhushan/plant-disease-recognition-ai-challenger> (accessed on 24 June 2026).
40. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
41. Ahmed, S.; Hasan, M.B.; Ahmed, T.; Sony, M.R.K.; Kabir, M.H. Less Is More: Lighter and Faster Deep Neural Architecture for Tomato Leaf Disease Classification. *IEEE Access* **2022**, *10*, 68868–68884. [[CrossRef](#)]
42. Agarwal, M.; Gupta, S.K.; Biswas, K.K. Development of Efficient CNN Model for Tomato Crop Disease Identification. *Sustain. Comput. Inform. Syst.* **2020**, *28*, 100407. [[CrossRef](#)]
43. Abbas, A.; Jain, S.; Gour, M.; Vankudothu, S. Tomato Plant Disease Detection Using Transfer Learning with C-GAN Synthetic Images. *Comput. Electron. Agric.* **2021**, *187*, 106279. [[CrossRef](#)]
44. Yang, S.; Jin, Y.; Lei, J.; Zhang, S. Multi-directional guidance network for fine-grained visual classification. *Vis. Comput.* **2024**, *40*, 8113–8124. [[CrossRef](#)]
45. Tao, J.; Li, X.; He, Y.; Islam, M.A. CEFW-YOLO: A High-Precision Model for Plant Leaf Disease Detection in Natural Environments. *Agriculture* **2025**, *15*, 833. [[CrossRef](#)]
46. Arshad, Z.; Javed, A.; Saudagar, A.K.J. ConvGeM-next: A deep learning framework for plant disease detection. *Front. Plant Sci.* **2026**, *17*, 1763739. [[CrossRef](#)] [[PubMed](#)]
47. Sumith, N.; Rashmi, M. A systematic review of deep learning and super resolution techniques for leaf level and canopy level plant disease detection. *Discov. Artif. Intell.* **2026**, *6*, 316. [[CrossRef](#)]
48. Yang, S.; Ren, X.; Zha, B.; Bao, Q.; Feng, B.; Xu, Z. EFSRNet: Multi-scale exposure normalization and dual-branch aggregation for overexposed face super-resolution. *Knowl.-Based Syst.* **2026**, *339*, 115578. [[CrossRef](#)]
49. Song, Z.; Zhu, Y.; Wang, D.; Liu, H.; Jiang, L.; Duan, Y.; Zhang, Z.; Li, S.; Li, J. TCLeaf-Net: A Transformer-Convolution Framework with Global-Local Attention for Robust In-Field Lesion-Level Plant Leaf Disease Detection. *arXiv* **2025**, arXiv:2512.12357. [[CrossRef](#)]
50. Xiao, W.; Shang, J.; Li, F.; Ao, O.; Wang, X.; Tian, S. Research on citrus leaf disease recognition using class-agnostic contrastive learning and supervised organizational mapping. *IEEE Access* **2025**, *13*, 101592–101608. [[CrossRef](#)]

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.