Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.082431

ARTICLE

Scale-Robust Cross-Scale Representation Learning for Aerial Crop Pest Recognition

Kemeng Zhu[1], Dingju Zhu[1,2,*], Shihua Mao[1], Jinchen Wu[3], Depeng Kong[4], Kaileung Yung[5] and Andrew W. H. Ip[6]

[1]School of Artificial Intelligence, South China Normal University, Foshan, China
[2]School of Computer Science, South China Normal University, Guangzhou, China
[3]Guangzhou Vocational College of Technology & Business, Guangzhou University, Guangzhou, China
[4]Department of McCormick School of Engineering, Northwestern University, Evanston, IL, USA
[5]Department of Industrial and Systems Engineering, Hong Kong Polytechnic University, Hong Kong, China
[6]Department of Mechanical Engineering, University of Saskatchewan, Saskatoon, SK, Canada

*Corresponding Author: Dingju Zhu. Email: zhudingju@m.scnu.edu.cn

Received: 16 March 2026; Accepted: 21 April 2026; Published: 15 June 2026

ABSTRACT: Unmanned aerial vehicles (UAVs) have become an increasingly important platform for agricultural remote sensing, yet the accurate recognition of pests and diseases is frequently compromised by drastic scale variability and complex environmental backgrounds. To address these challenges, this study introduces a novel attention-driven approach centered on a Multi-Scale Grouped Channel–Spatial Dual Attention (MS-GCDA) mechanism. The MS-GCDA module achieves robust feature calibration by decoupling and jointly modeling multi-scale spatial contexts and grouped channel dependencies, which significantly enhances the model's sensitivity to fine-grained disease symptoms while suppressing background clutter. This core mechanism is integrated into Augmented EfficientNet (AugEffNet), a lightweight architecture further optimized for edge deployment through a learnable Latent Feature Projection (LFP) strategy that bridges the gap between high-dimensional feature extraction and real-time inference. Experimental evaluations on a self-collected luffa dataset and the public AppleLeaf9 benchmark demonstrate the superior efficacy of the proposed MS-GCDA mechanism. The framework achieves an accuracy of 97.27% and an F1-score of 97.05%, outperforming representative lightweight models. Notably, while the MS-GCDA ensures diagnostic precision under varying flight altitudes, the integration of LFP optimizes the deployment pipeline, resulting in a 94.7% reduction in inference latency (from 127.57 to 6.68 ms) and a decrease in computational cost to 122.88M Floating Point Operations per Second (FLOPs). Real-world field experiments validate that the synergy between scale-aware attention and efficient feature projection provides a reliable and scalable solution for autonomous plant protection in precision agriculture.

KEYWORDS: Lightweight network; unmanned aerial vehicles; complex context; plant disease; image recognition

1 Introduction

Unmanned aerial vehicles (UAVs) have become an increasingly important platform for agricultural remote sensing, enabling timely and fine-grained field monitoring through flexible deployment and high spatial resolution imaging [1]. In precision agriculture, early detection of crop pests and diseases is critical to support targeted intervention strategies, reducing excessive pesticide application, and mitigating yield losses. Compared with satellite-based systems, UAV platforms enable more frequent field-scale observations,

making them particularly suitable for monitoring heterogeneous crop health conditions in real-world farming environments.

Despite these advantages, reliable pest and disease recognition from UAV imagery remains a challenge. Unlike ground-level imaging, UAV-acquired data must overcome pronounced scale variability—driven by altitude and perspective changes—and complex background clutter, such as soil, shadows and overlapping vegetation [2]. These environmental factors demand feature representations that are simultaneously robust to resolution fluctuations and resilient to visual interference. From a deployment perspective, the limited computational and energy budgets of onboard edge devices further restrict the use of heavy neural networks [3]. Consequently, a practical UAV-oriented model must strike a delicate balance between scale robustness, background resilience, and real-time inference efficiency.

Recent advances in deep learning have significantly improved automated plant recognition. While early Convolutional neural network (CNN)-based models achieved high accuracy under controlled conditions [4], recent research has shifted toward optimizing UAV operations and recognition algorithms. For instance, bio-inspired meta-heuristic algorithms, such as dragonfly and firefly algorithms, have been effectively applied to UAV path scheduling and target trajectory optimization to enhance operational efficiency [5]. For aerial visual understanding, lightweight transformer architectures with adaptive rotational convolutions have shown outstanding performance in object detection from remote sensing imagery [6] demonstrating the value of rotation-aware and scale-adaptive designs for UAV scenarios. In terms of architectural efficiency, models like YOLO-PDGT have demonstrated the potential of lightweight designs for specific tasks such as unripe pomegranate detection and counting [7]. Furthermore, traditional image processing techniques continue to provide foundational insights for diagnosing specific pathologies, such as Alternaria disease and Leafminer pests on tomato leaves [8].

To bridge these developments, lightweight architectures and attention mechanisms have emerged as promising solutions to improve the accuracy-efficiency trade-off [9–12]. Attention mechanisms, in particular, enhance discriminative learning by emphasizing disease-relevant regions [13,14]. However, most existing UAV-oriented approaches either rely on single-scale representations or are trained primarily on close-range data, failing to provide a unified solution that addresses scale, background, and deployment constraints simultaneously.

Uniquely, our approach bridges the gap between accessible close-range training data and challenging aerial deployment by integrating a scale-aware dual-attention mechanism with a downstream dimensionality-reduction strategy, ensuring robust recognition without the need for exhaustive UAV-specific datasets.

Motivated by these challenges, this study proposes a lightweight UAV-oriented framework. To address the requirements of scale robustness, background resilience, and efficiency, we integrate two strategies: (i) a Multi-Scale Grouped Channel–Spatial Dual Attention (MS-GCDA) module for robust feature learning, and (ii) a deployment-stage feature compression strategy based on Latent Feature Projection (LFP) to minimize inference latency on edge devices.

The main contributions of this study are summarized as follows:

- We develop a unified lightweight framework that systematically addresses scale variability, background interference, and hardware constraints through integrated architectural and deployment-stage design.
- We design a novel MS-GCDA module that jointly models spatial and channel-wise dependencies across multiple scales, enabling invariant feature learning under significant UAV-induced scale fluctuations.
- We introduce a LFP-based feature projection strategy that substantially accelerates inference while maintaining accuracy, making the framework viable for real-time edge deployment.

- We validate the framework on both self-collected loofah and public datasets, demonstrating superior performance and real-world applicability through actual UAV field experiments.

The paper is structured as follows: Section 2 describes the Data and Methods, detailing the data acquisition process, the proposed MS-GCDA module, and the overall architectural design. Section 3 presents the experimental Results, including performance metrics and a comparative analysis with state-of-the-art models. Section 4 provides in-depth Discussions, focusing on ablation studies and visualization analysis to validate the effectiveness of the proposed components. Finally, Section 5 summarizes the key findings of this study and outlines future research directions in the Conclusions.

2 Data and Methods

2.1 Datasets

The loofah images used in this study were sourced from the Nanfeng base of Guangdong Zhaoqing Fengyue Agricultural Development Co., Ltd., Zhaoqing City, Guangdong Province, China (23°44′ N, 111°47′ E). The image acquisition area is shown in Fig. 1. Data collection was conducted between 11 July and 14 July 2024, specifically during 9–11 a.m. and 3–5 p.m. to capture a wide spectrum of natural lighting conditions and shadow angles.

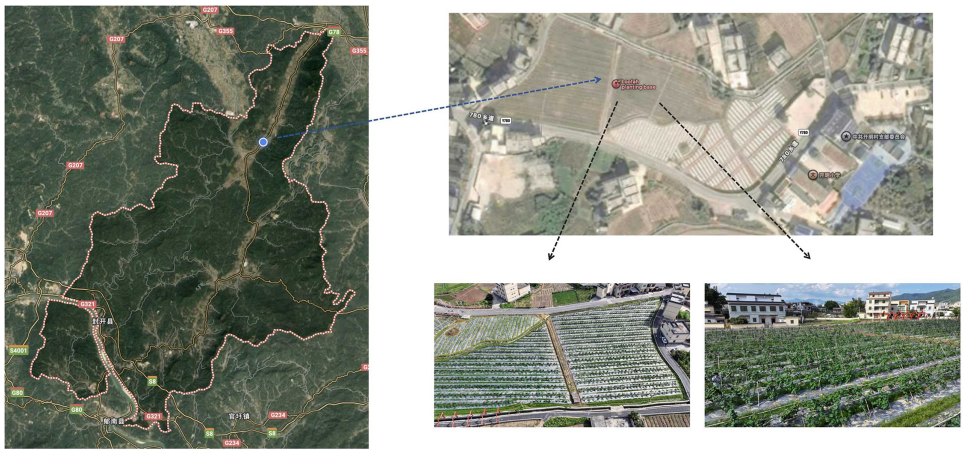

Figure 1: Image acquisition area.

To ensure the robustness of the recognition model across varying hardware, images were captured using a combination of professional and consumer-grade equipment: a Canon EOS R6 Mark II (full-frame mirrorless, approx. 24.2 MP) and various representative smartphones (e.g., iPhone 14 Pro Max, Xiaomi 12S, and OPPO Reno4 SE) with resolutions ranging from 12 to 50 MP. This diversity in imaging sensors and optical qualities simulates the heterogeneous data sources typically encountered in practical UAV-based agricultural monitoring. A total of 2,929 loofah images were collected, categorized into: downy mildew (475), diaphania indica (550), healthy (732), liriomyza (496), and needle peak (676), as illustrated in Fig. 2.

Two distinct datasets were employed in this study to provide a comprehensive evaluation of the proposed framework. The self-collected Loofah dataset focuses on verifying the model's performance under authentic UAV remote sensing conditions, characterized by complex field backgrounds and varying flight altitudes. In contrast, the AppleLeaf9 public dataset, containing a larger volume of standardized samples, was utilized as a benchmark to validate the model's generalization capability across different crop species and diverse pathological manifestations.

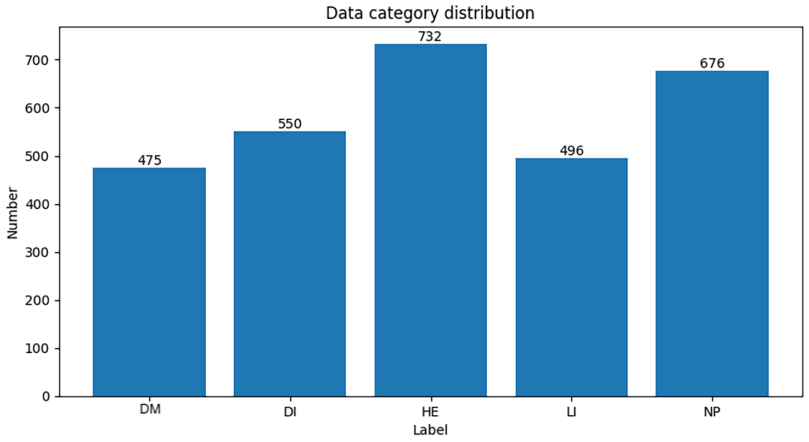

Figure 2: Data category distribution of loofah.

Some images of loofahs are shown in Fig. 3. The figure shows sample images for each category, corresponding to the following labels: Downy Mildew (DM), Diaphania Indica (DI), Healthy (HE), Liriomyza (LI), and Needle Peak (NP). The shapes of the loofahs in the images differ, the image backgrounds are complex, the lighting is uneven, and the severity of pests and diseases varies. Identifying loofah pests and diseases in field settings is a substantial challenge.

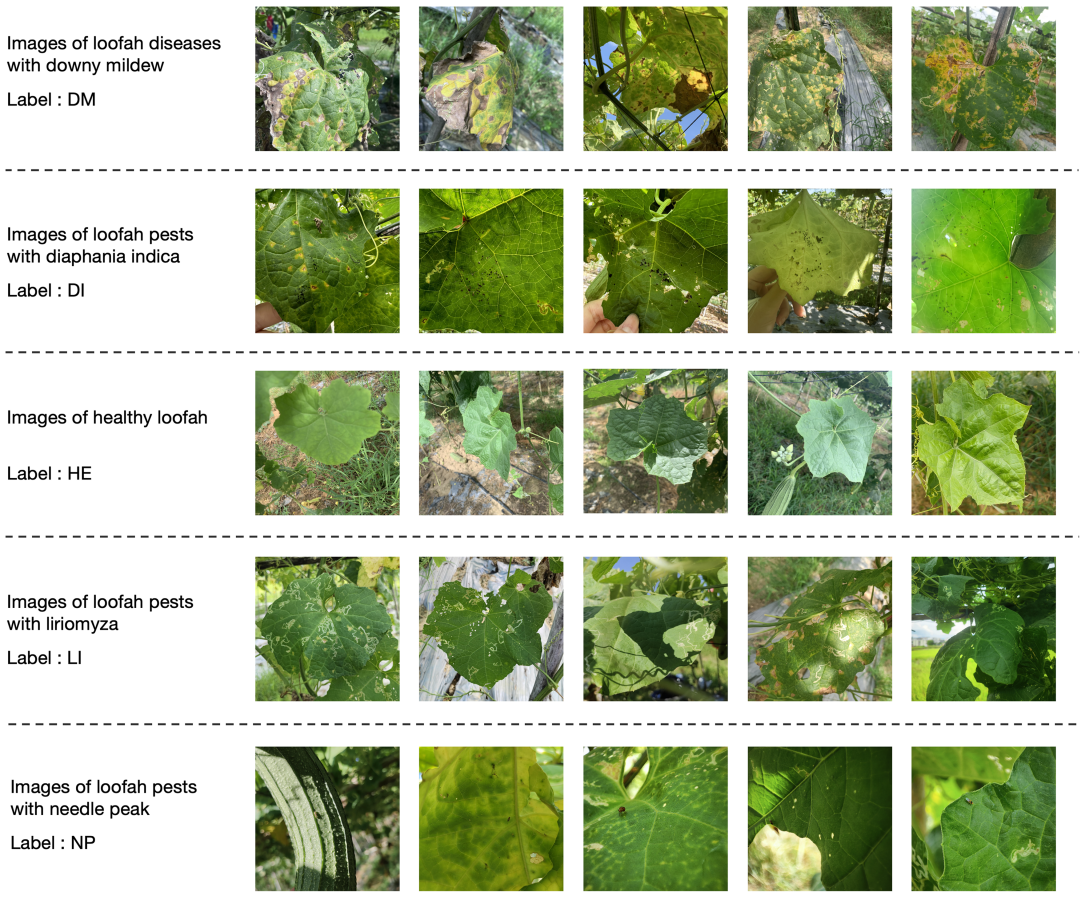

Figure 3: Some images of loofah.

Table 1 shows the corresponding symptoms. The input images were resized to 224 × 224, and the dataset was divided into training, validation, and test sets with a 6:2:2 ratio.

Table 1: Loofah pests and diseases symptoms by category.

Category	Main Symptom
Downy Mildew	Irregular yellow spots turning brown; purplish-black mold on the underside in humidity
Diaphania indica	Larvae skeletonize leaves, initially leaving veins, later consuming entire leaves
Healthy	Leaves are deep green, thick, and intact with a smooth surface and distinct veins
Liriomyza	Larvae create transparent, winding tunnels (mines) inside the leaf tissue
Needle Peak	Small brown spots enlarge into circular or elliptical lesions with needle-like protrusions

The dataset used in this experiment is AppleLeaf9, a collection of images of apple diseases published by Yang et al. [9]. As shown in Fig. 4, it comprises 14,582 images. The dataset was created by combining apple leaf images from four different datasets: PlantVillage, ATLDSD [15], PPCD2020 [16], and PPCD2021 [17]. The AppleLeaf9 dataset consists of images of healthy apple leaves and eight types of apple disease leaves. 94% of the images were captured in complex natural environments in the field, with varying shooting angles, lighting conditions, and backgrounds. Fig. 5 displays the apple leaf images for each category in the dataset, while Table 2 shows the corresponding symptoms. The input images were resized to 224 × 224, and the dataset was split into training, validation, and test sets in a 6:2:2 ratio.

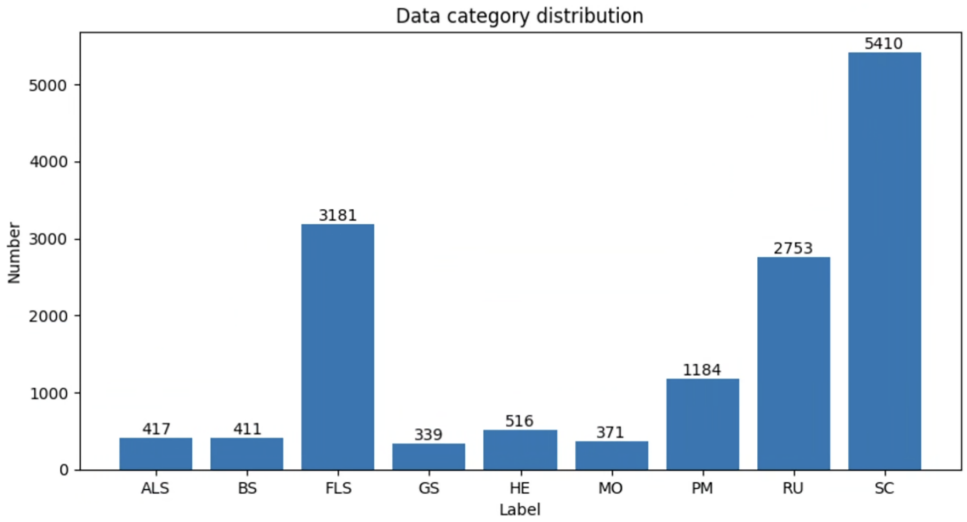

Figure 4: Data category distribution of AppleLeaf9.

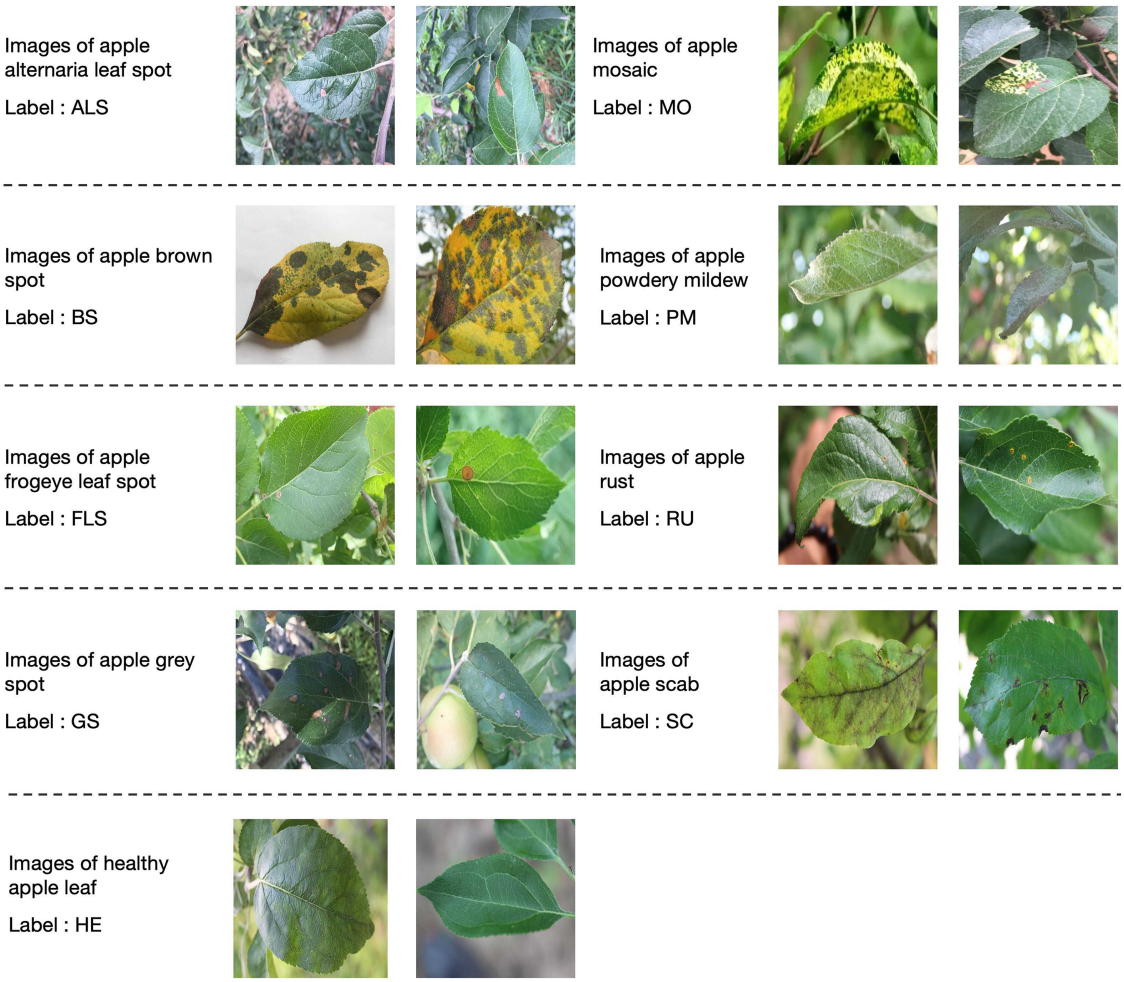

Figure 5: Some images of AppleLeaf9.

Table 2: Apple diseases symptoms by category.

Category	Main Symptom
Alternaria leaf spot	Spots usually have small, round brown or black lesions
Brown spot	Has distinctive dark brown spots
Frogeye leaf spot	Spots turn brown in the center, dark brown to purple at the edges
Grey spot	Early stages show subrounded yellowish-brown lesions that turn gray with time
Healthy	Characterized primarily by green foliage
Mosaic	Bright yellow spots all over the leaves
Powdery mildew	The leaves have many distinct white spots
Rust	Spots are typically rusty yellow with brown pinpoints at the center
Scab	Spots are velvety with fringed margins

2.2 Multi-Scale Grouped Channel–Spatial Dual Attention Module (MS-GCDA)

UAV-based crop pest and disease recognition poses distinct challenges compared with conventional ground-level image classification. Due to variations in flight altitude, camera viewing angles, and scene composition, target objects in UAV imagery often exhibit pronounced scale variability, while complex backgrounds such as soil, shadows, and overlapping vegetation introduce substantial interference. These factors make it difficult for standard convolutional neural networks to consistently capture discriminative features across different spatial scales and environmental contexts.

Existing lightweight recognition models typically rely on single-scale feature representations or employ channel- or spatial-attention mechanisms independently. Although such approaches can improve performance to some extent, they often struggle to address cross-scale feature misalignment and background clutter simultaneously, particularly under the computational constraints of UAV edge deployment. In practical remote sensing scenarios, an effective recognition model should be able to learn scale-robust representations that adaptively emphasize informative spatial regions across multiple resolutions while selectively enhancing semantically relevant channels.

Motivated by these considerations, we introduce a multi-scale grouped channel–spatial dual attention (MS-GCDA) module as the core of our cross-scale representation learning paradigm. The proposed module is designed to jointly model spatial and channel-wise dependencies through a sequential refinement process, thereby improving robustness to variations in object size and complex background interference.

Specifically, the MS-GCDA module implements a "detect-then-rectify" strategy. It first employs a scale-adaptive structural sampler to capture contextual information at different resolutions, which is particularly important for accommodating scale changes induced by UAV imaging geometry. Following this, a semantic importance rectifier is employed to model inter-channel relationships in a computationally efficient manner, allowing the network to emphasize disease-relevant semantic features while suppressing background noise. By integrating spatial and channel attention in this sequential framework, MS-GCDA enables more discriminative and stable feature learning under real-world UAV remote sensing conditions.

The design of the MS-GCDA module is rooted in this sequential calibration logic, drawing conceptual insights from prior studies on multi-scale feature fusion and channel attention while being specifically adapted to the challenges of UAV-based remote sensing. In particular, the dynamic multi-scale fusion mechanism of SKNet [18] informs the multi-scale spatial modeling strategy, the hierarchical grouped channel representation of Res2Net [19] motivates the grouped channel interaction design, and the channel attention modeling of Squeeze-and-Excitation Networks [20] provides a foundation for enhancing semantic feature discrimination. Building upon these ideas, the proposed MS-GCDA module integrates spatial and channel attention within a unified framework tailored to address scale variability and background interference in UAV imagery, as illustrated in Fig. 6.

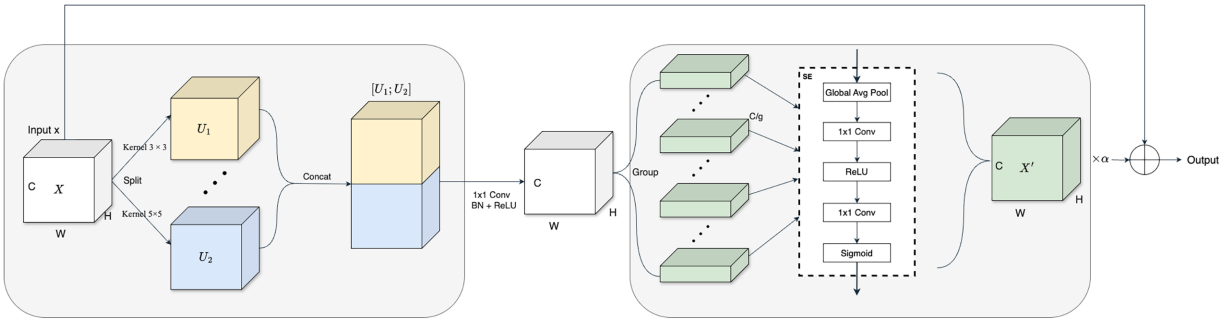

Figure 6: MS-GCDA module.

2.2.1 Sequential Scale-Aware Calibration

The primary objective of the MS-GCDA module is to enhance recognition performance by explicitly modeling the "where" (spatial saliency) and "what" (semantic relevance) of pest-related features. As illustrated in Fig. 6, MS-GCDA is formulated as a sequential multi-stage refinement process. Unlike conventional parallel attention mechanisms, it implements a "detect-then-rectify" strategy: first extracting multi-scale spatial contexts to localize informative regions across varying altitudes, and then applying grouped channel attention to recalibrate semantic importance. This synergistic design addresses the intra-class variance and background noise inherent in UAV-based agricultural monitoring.

2.2.2 Multi-Scale Spatial Attention Branch

The initial stage of MS-GCDA serves as a Scale-Adaptive Structural Sampler. To address the variance in object resolution caused by UAV altitude fluctuations, we construct a parallel mapping function to capture features across different receptive fields:

$$X_k = \mathrm{DWConv}_k(X), \quad \forall k \in \{3, 5\} \tag{1}$$

Intuitively, Eq. (1) allows the model to perceive the crop from multiple granularities simultaneously; while the 3 × 3 kernel identifies fine-grained pest textures, the 5 × 5 kernel captures the broader topology of disease lesions. This ensures feature invariance across different imaging scales.

The choice of kernel sizes $\{3, 5\}$ is a strategic trade-off for UAV edge deployment: 3 × 3 kernels provide high-precision local feature extraction, while 5 × 5 kernels offer a sufficient contextual envelope for larger symptoms. Larger kernels (e.g., 7 × 7) were excluded to prevent an exponential increase in computational overhead on resource-constrained devices.

These multi-scale features are subsequently aggregated via concatenation to form a unified spatial representation:

$$X_{\mathrm{cat}} = X_3 \parallel X_5 \tag{2}$$

where $\parallel$ denotes the concatenation operation. To distill these signals, a 1 × 1 convolution is applied to generate the spatially-refined feature map X_f:

$$X_f = \delta(\mathrm{BN}(W_p * X_{\mathrm{cat}})) \tag{3}$$

Conceptually, Eq. (3) acts as a spatial information bottleneck that projects high-dimensional multi-scale features into a compact saliency map, effectively suppressing background noise such as soil reflectance

and leaf shadows. The core contribution of this branch is the establishment of scale-robustness in complex field environments.

2.2.3 Grouped Channel Attention Branch

The spatially-refined feature X_f is then passed to the Semantic Importance Rectifier. To maximize efficiency under edge constraints, we partition X_f into G disjoint groups $[X_f^{(1)}, \ldots, X_f^{(G)}]$, each undergoing an independent gating process:

$$A^{(g)} = \sigma\left(W_2^{(g)} \cdot \delta\left(W_1^{(g)} \cdot \mathrm{GAP}\left(X_f^{(g)}\right)\right)\right) \tag{4}$$

Eq. (4) functions as a software-defined filter that computes a "relevance score" for each channel group, magnifying channels that carry critical diagnostic markers while dimming those representing redundant background data. The contribution of this grouped mechanism lies in enhancing semantic discriminability with minimal computational overhead.

The rectified features for each group are obtained via Hadamard product:

$$\tilde{X}_f^{(g)} = A^{(g)} \otimes X_f^{(g)} \tag{5}$$

and the final attention-enhanced feature $\tilde{X}_f$ is reconstructed by concatenating all group-wise rectified features:

$$\tilde{X}_f = \mathrm{Concat}\left(\tilde{X}_f^{(1)}, \tilde{X}_f^{(2)}, \ldots, \tilde{X}_f^{(G)}\right) \tag{6}$$

2.2.4 Identity-Preserving Residual Fusion

To facilitate stable gradient flow and prevent the loss of raw textural details, the final output Y is defined through an Adaptive Residual Integration:

$$Y = X + \alpha \cdot \tilde{X}_f \tag{7}$$

Eq. (7) represents a residual learning objective where the original input X maintains structural identity, and the learned attention $\tilde{X}_f$ provides diagnostic enhancements.

In summary, the contributions of MS-GCDA to performance enhancement are three-fold: (1) The multi-scale DWConv layers provide scale-invariance against flight height changes; (2) The bottleneck projection contributes to background suppression; and (3) The grouped attention offers efficient semantic calibration. This integration ensures the model achieves high accuracy while remaining optimized for real-time UAV deployment.

2.3 Efficient Feature Modeling Network Based on EfficientNet (AugEffNet)

EfficientNet is widely recognized for its favorable balance between recognition accuracy and computational efficiency, making it an ideal foundation for resource-constrained UAV remote sensing applications [21]. In aerial monitoring scenarios, the backbone must not only be lightweight but also capable of generating high-quality feature hierarchies that support scale-robust representation learning. Accordingly, this study adapts the EfficientNet-B0 architecture to construct AugEffNet, which serves as the primary feature extraction engine for our framework.

The fundamental building block of AugEffNet is the Attention-enhanced Mobile Inverted Bottleneck Convolution Block with Expansion Ratio 6 (AttMBConv6) module, an attention-enhanced version of the standard mobile inverted bottleneck convolution, as illustrated in Fig. 7. While retaining the efficiency of depthwise separable convolutions and inverted residuals, AttMBConv6 incorporates local channel-wise recalibration. This internal refinement ensures that as the network processes information, it can adaptively emphasize diagnostic lesion patterns while suppressing background noise. By embedding this mechanism directly within the MBConv stages, the backbone performs preliminary semantic filtering before the features reach the cross-scale refinement stage.

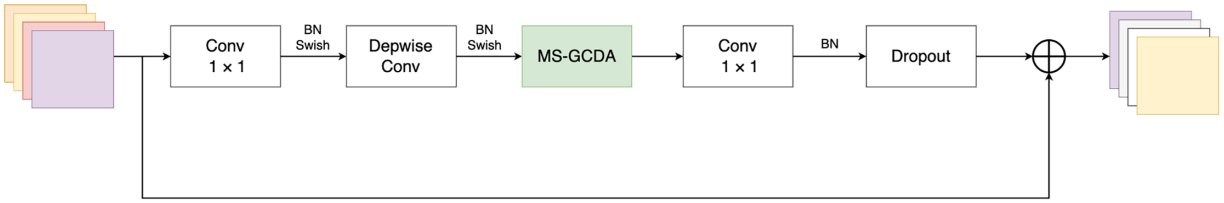

Figure 7: AttMBConv6 module.

Based on these blocks, the overall AugEffNet architecture is structured to provide the multi-level granularity required for aerial vision, as shown in Fig. 8. The network follows a stage-wise design, stacking AttMBConv6 blocks with progressively increasing channel dimensions and decreasing spatial resolutions. To optimize the trade-off between representational depth and edge-deployment latency, we selectively integrate the attention-enhanced blocks at the 10th and 15th stages of the architecture. This strategic placement ensures that high-level semantic abstractions are sufficiently refined without incurring prohibitive computational overhead. The AugEffNet architecture is organized into eight distinct stages based on feature resolution. To support cross-scale representation learning, intermediate feature maps are extracted from Stage 4, Stage 6, and Stage 8, representing high, medium, and low spatial resolutions, respectively.

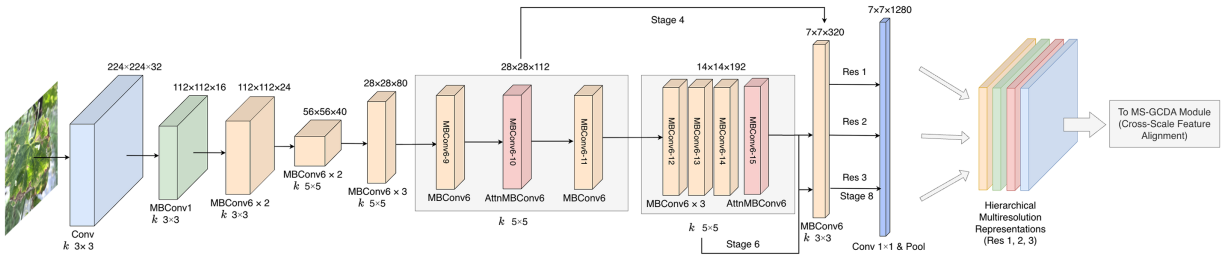

Figure 8: AugEffNet model.

Crucially, AugEffNet is configured to produce intermediate feature maps from multiple depths. These hierarchical representations capture complementary spatial details and semantic cues, serving as the essential inputs for the subsequent MS-GCDA module. By preserving the lightweight essence of EfficientNet while introducing stage-wise feature modeling, AugEffNet establishes a stable and robust foundation for reconciling scale variability and background heterogeneity in UAV imagery. This architectural design ensures that the framework remains well-suited for real-time deployment on UAV edge devices.

2.4 Deployment-Stage Feature Projection and Classification

To bridge the gap between high-dimensional feature extraction and real-time inference on resource-constrained UAV edge platforms, we implement a Latent Feature Projection (LFP) strategy. While conventional lightweight models often directly append a global average pooling layer to a large softmax classifier, this can lead to significant computational redundancy and noise sensitivity in unstructured field environments.

As illustrated in our architectural design, the features refined by the MS-GCDA module are fed into a Learnable Linear Bottleneck. Unlike static dimensionality reduction techniques such as principal component analysis, our LFP head is optimized via backpropagation, allowing the model to adaptively project complex multi-scale features into a task-specific latent space. This process is formulated as:

$$Z = \delta(\mathrm{BN}(W_{proj} \cdot f_{gap}(X) + b_{proj})) \tag{8}$$

where $f_{gap}(\cdot)$ denotes global average pooling, and W_{proj} represents the learnable projection matrix that compresses the feature dimensions to a compact 128-dimensional vector. This bottleneck layer functions as a structural regularizer, distilling essential diagnostic markers while discarding redundant environmental information.

The final classification is performed by a linear head optimized for the latent manifold:

$$\hat{y} = \mathrm{Softmax}(W_{cls} \cdot Z + b_{cls}) \tag{9}$$

By integrating this learnable projection, the framework achieves a dual benefit: it significantly reduces the number of parameters in the final fully-connected layer and enhances the discriminative stability of the model under variable UAV imaging conditions. As demonstrated in our benchmark experiments (see Section 3.5), this "LFP" approach yields superior F1-scores compared to traditional non-learnable classifiers, while maintaining an extremely low inference latency of 6.68 ms on edge hardware.

2.5 UAV Edge Device

Field image data were acquired using a DJI Mini 4 Pro consumer-grade unmanned aerial vehicle (UAV) platform, as shown in Fig. 9. Equipped with a 1/1.3-inch complementary metal-oxide-semiconductor sensor, the UAV can capture video at up to 4K/60fps and still images at 48 megapixels, providing high-resolution raw data for disease identification. During field operations, the UAV performed low-speed cruising at an altitude of 3–5 m above the loofah field. Image and video data of the crop canopy were acquired from a nadir (vertical) or moderately oblique viewing angle to ensure image clarity and the visibility of disease symptoms.

2.6 Experimental Environment and Evaluation Indicators

The experiment was conducted on Windows 10 using the PyTorch framework for model training and testing. The graphics card used was an NVIDIA RTX 4060 Ti with 16 GB of video memory. Python version 3.9.21, PyTorch version 2.7.0, and Compute Unified Device Architecture (CUDA) Application Programming Interface (API) version 12.6 was utilised in this experiment. version 12.6 were utilised in this experiment. The model was optimised using the Adaptive Moment Estimation (Adam) algorithm with a learning rate of 0.0005. For the loofah dataset, training was conducted with a batch size of 64 and the cross-entropy loss function. To prevent overfitting, we implemented an early stopping strategy that monitored the validation loss with a patience of 20 epochs, setting the maximum number of epochs to 100.

Figure 9: Drone equipment.

The experiments utilised various evaluation metrics to comprehensively assess the model, including Accuracy, Recall, Precision, and F1-Score. Accuracy measures the proportion of correctly classified samples out of the total number of samples. Recall measures the proportion of all positive samples that are correctly identified as positive. Precision measures the proportion of samples with positive classification results that are actually positive. F1-Score is a weighted average of Recall and Precision. Eqs. (10)–(13) show the calculation process for each metric:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$F1\text{-}Score = \frac{2TP}{2TP + FP + FN} \tag{13}$$

where TP is the number of positive samples correctly identified as positive, FP is the number of negative samples incorrectly identified as positive, FN is the number of positive samples incorrectly identified as negative, and TN is the number of negative samples correctly identified as negative. The Params, the Floating Point Operations (FLOPs), and the Test time were used in the experiments as metrics for evaluating the model's lightness.

3 Results

3.1 Identification Performance of AugEffNet

Fig. 10 shows the accuracy and loss values of the proposed AugEffNet for training and testing datasets in 80 epochs. The model's loss value drops rapidly during training and stabilizes after approximately 10 epochs,

indicating that it has converged in a short time. In terms of accuracy, the model rapidly improved to over 90% in the initial stage and maintained it at around 95% in the later stage, demonstrating strong fitting ability.

To further evaluate the model's classification performance, this paper uses a confusion matrix for analysis, and the results are shown in Fig. 11. Fig. 11a shows the unnormalized confusion matrix, and Fig. 11b shows the normalized result. It can be seen from the figure that:

- The model achieves high classification accuracy across most categories.
- The recognition accuracy rates of HE and NP classes reached 99.32% and 98.53%, respectively, indicating that the model has a strong discriminative ability for obvious feature classes.
- For the LI class, while achieving a slightly lower accuracy of approximately 89%, it still maintains a competitively high recognition rate.
- Overall, the model demonstrates excellent classification performance, especially maintaining robust performance in scenarios with uneven data distribution.

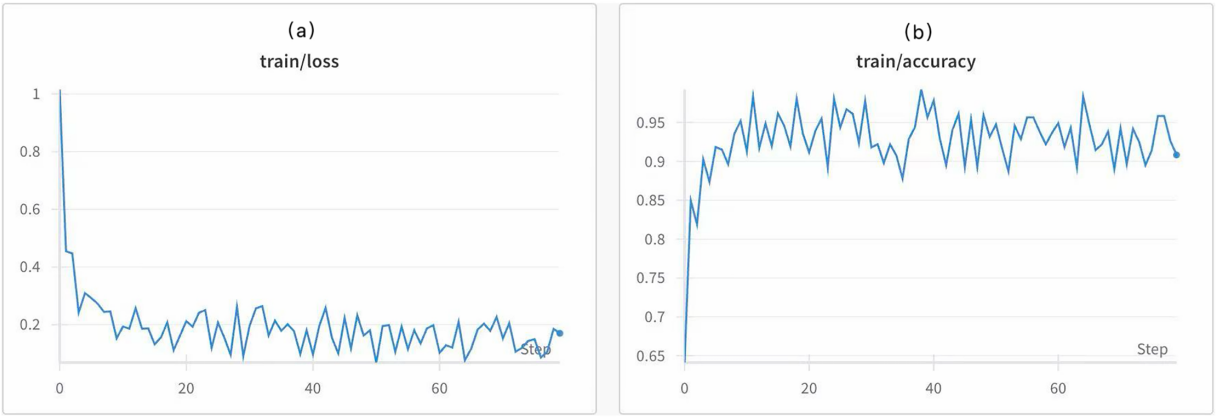

Figure 10: Convergence comparison: (**a**) loss; (**b**) accuracy.

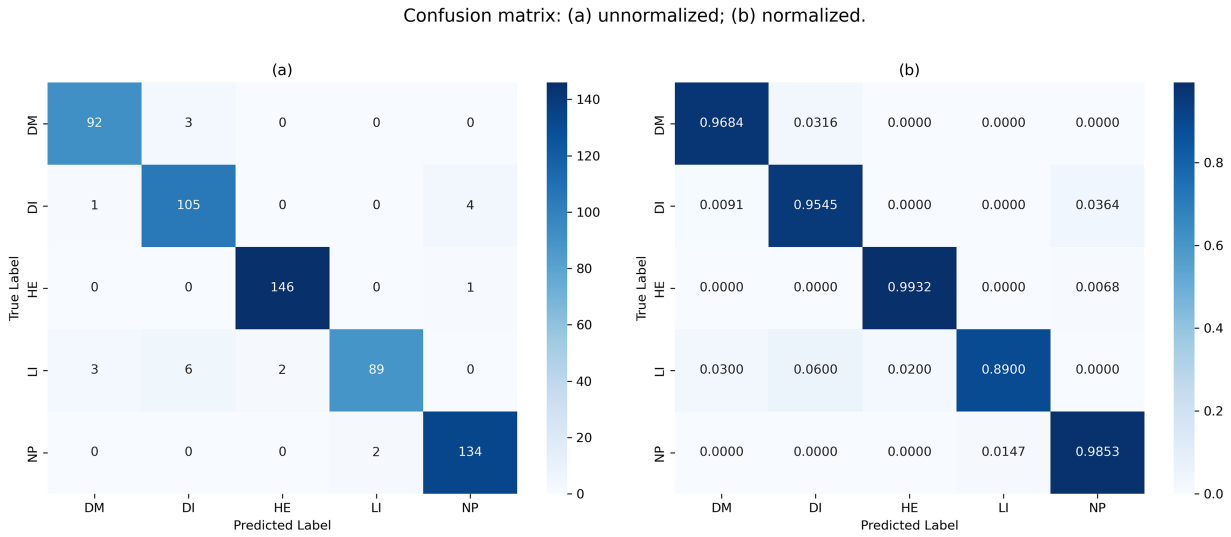

Figure 11: Confusion matrix: (**a**) unnormalized; (**b**) normalized.

The confusion matrix shows that AugEffNet achieves high recognition accuracy across various loofah pests and diseases, with no serious recognition bias.

3.2 Comparison Experiment

To validate that the AugEffNet model can maintain both low computational complexity and outstanding disease recognition performance on the Loofah dataset, we compared and analyzed it against lightweight models including EfficientNet-B0 [21], ShuffleNetV2 [22], DenseNet121 [23], MobileNetV3-small [24], and GhostNetV2 [25]. As shown in Table 3, AugEffNet achieved a significantly leading accuracy of 96.26% on the test set (a 0.85 percentage point improvement over the second-best model, ShuffleNetV2, at 95.75%), while also attaining the highest scores in F1-Score (96.28%), Recall (96.23%), and Precision (96.53%).

Table 3: Identification performance comparison among different CNN models.

Model	Accuracy (%)	F1-Score (%)	Recall (%)	Precision (%)	FLOPS (M)	Params (M)	Test time (ms)
EfficientNet_b0	95.41	95.00	94.85	95.28	27.02	1.048	115.55
ShuffleNetV2	95.75	95.26	95.15	95.12	151.69	1.259	68.88
DenseNet121	94.73	94.21	94.14	94.58	2895.99	6.959	190.76
MobileNetV3_small	92.69	91.90	91.63	92.61	61.46	1.523	53.13
GhostNetV2	95.74	95.27	95.18	95.58	183.81	4.880	82.54
AugEffNet (Ours)	96.60	96.28	96.23	96.53	161.22	0.457	127.57

In terms of lightweight performance, AugEffNet requires only 0.457M parameters, which is significantly lower than those of the other compared models. In addition, its computational load (161.22M FLOPs) is less than 5.6% of that of DenseNet121.

A comprehensive performance analysis further highlights the core advantages of AugEffNet: Compared to the similarly lightweight MobileNetV3-small-100 (92.69% accuracy), it achieves a substantial 4.21% accuracy improvement. Compared to the computationally intensive DenseNet121 (94.73% accuracy), AugEffNet achieves 96.60% accuracy while using only 5.57% of the parameters.

To validate AugEffNet's generalization ability, we evaluated it on the public benchmark dataset AppleLeaf9, ensuring fair and reproducible comparisons with existing methods. The model was compared and analyzed against lightweight models including EfficientNet-B0 [21], GhostNetV2 [25], MobileNetV3-small [24], FasterNet-T0 [26] and ShuffleNetV2 [22]. As shown in Table 4, AugEffNet achieved a remarkable accuracy of 97.33% and an F1-Score of 96.72%, significantly outperforming all compared lightweight models. The F1-Score shows an improvement of 1.52 percentage points over the second-best model (MobileNetV3-small-50 at 95.20%). Meanwhile, it also achieved the highest recorded scores in both Recall (96.32%) and Precision (97.18%), demonstrating the model's robustness in fine-grained classification tasks.

Table 4: Recognition results of different lightweight models on AppleLeaf9 dataset.

Model	Accuracy (%)	F1-Score (%)	Recall (%)	Precision (%)	FLOPS (M)	Params (M)	Test time (ms)
EfficientNet_b0	97.02	95.13	94.83	95.48	27.03	1.054	86.56
GhostNetV2	96.88	95.19	95.22	95.70	183.80	4.880	54.57

(Continued)

Table 4 (continued)

Model	Accuracy (%)	F1-Score (%)	Recall (%)	Precision (%)	FLOPS (M)	Params (M)	Test time (ms)
MobileNetV3_small	96.60	95.20	95.00	95.46	55.51	1.560	93.30
FasterNet_t0	96.81	94.74	94.41	95.14	339.38	2.640	102.20
ShuffleNetV2	97.19	94.80	94.10	95.54	151.71	1.280	94.20
AugEffNet (Ours)	97.33	96.72	96.32	97.18	161.22	0.462	82.29

3.3 Comparison with State-of-the-Art (SOTA) Models

To comprehensively evaluate the performance of AugEffNet, we compared it with several recent state-of-the-art models optimized for edge devices, including ConvNeXt-small [27], EdgeNeXt-xx-small [28], YOLOv8-nano [29], TinyViT-5m [30], and MobileViTv2-1.0 [24]. These models represent a diverse range of architectural designs, from optimized convolutional neural networks to lightweight vision transformers. The comparative results, including recognition accuracy, F1-score, computational complexity (FLOPs), parameter count, and actual test time, are summarized in Table 5.

Table 5: Identification performance comparison with recent SOTA models on edge devices.

Model	Accuracy (%)	F1-Score (%)	Recall (%)	Precision (%)	FLOPS (M)	Params (M)	Test time (ms)
ConvNeXt-small	95.80	95.42	95.31	95.54	1669.22	49.414	218.44
EdgeNeXt-xx-small	96.25	95.87	95.76	96.01	242.36	5.279	98.66
YOLOv8-nano	95.03	94.57	94.42	94.73	213.64	1.356	72.11
TinyViT-5m	94.37	93.86	93.71	94.02	258.77	5.460	76.33
MobileViTv2-1.0	95.90	95.53	95.50	95.45	370.15	6.366	105.29
AugEffNet (Ours)	96.60	96.28	96.23	96.53	161.22	0.457	127.57

The experimental results demonstrate that AugEffNet achieves the highest recognition performance while maintaining an ultra-lightweight structural footprint. Specifically, AugEffNet reaches an accuracy of 96.60% and an F1-score of 96.28%, outperforming EdgeNeXt-xx-small by 0.35% and ConvNeXt-small by 0.80%. It is noteworthy that ConvNeXt-small, despite having a significantly larger parameter size (49.414 M), yields a lower accuracy than our proposed model. This superiority highlights the effectiveness of the Multi-Scale Grouped Channel–Spatial Dual Attention (MS-GCDA) mechanism in distilling critical diagnostic features from complex agricultural backgrounds.

In terms of model efficiency, AugEffNet exhibits a decisive advantage in parameter economy and computational cost. With only 0.457M parameters, AugEffNet is more than ten times smaller than EdgeNeXt-xx-small (5.279M) and TinyViT-5m (5.460M). Furthermore, AugEffNet achieves the lowest computational intensity with 161.22M FLOPs, representing a substantial reduction compared to MobileViTv2-1.0 (370.15 M). Although the actual test time of AugEffNet (127.57 ms) is higher than that of YOLOv8-nano (72.11 ms), the significant gains in recognition precision and the dramatic reduction in memory footprint make AugEffNet a more balanced and practical solution for high-precision pest and disease monitoring on resource-constrained UAV platforms.

3.4 Ablation Experiments

To validate the effectiveness of each component in the proposed AugEffNet model, we conducted ablation experiments on the Loofah dataset and designed five model variants for comparative analysis. Variant 1 removes the MS-GCDA module from the building block. Variant 2 removes the linear classification head classifier and replaces it with a traditional fully connected layer. Variant 3 uses the CA module instead of the SE module. Variant 4 uses the Convolutional Block Attention Module (CBAM) module to replace the SE module. The last row presents the performance of the complete AugEffNet model.

As shown in Table 6, removing the MS-GCDA module resulted in the most significant performance drop, with accuracy decreasing from 96.60% to 95.49% and F1-score declining by 1.07%. This indicates that MS-GCDA plays a crucial role in enhancing multi-scale contextual feature representation for fine-grained disease recognition. It is worth noting that this variant achieved the lowest FLOPs and parameter count, but at the cost of recognition performance, reflecting a trade-off between efficiency and accuracy. Furthermore, removing the Linear Classification Head (LCH) also led to a slight performance degradation (F1-score decreased by 0.34%), suggesting the advantage of the lightweight classifier design in balancing performance and efficiency. Notably, although the FLOPs and parameter count are the same as the full model, this variant has a slightly shorter test time.

Table 6: Recognition results of different variants of AugEffNet on the loofah dataset.

Model	Accuracy (%)	F1-Score (%)	Recall (%)	Precision (%)	FLOPS (M)	Params (M)	Test time (ms)
Variant 1	95.49	95.21	94.85	95.18	27.02	0.458	125.35
Variant 2	96.26	95.94	95.83	96.16	161.22	0.457	114.72
Variant 3	95.75	95.30	95.00	95.86	107.99	0.119	135.30
Variant 4	95.58	95.16	95.08	95.50	31.82	0.130	128.69
AugEffNet	96.60	96.28	96.23	96.53	161.22	0.457	127.57

To evaluate the effect of different channel attention strategies, the SE module was replaced with CBAM and CA modules. Although both variants outperformed Variant 1, their performance did not match that of the original SE attention strategy. Variant 3 achieved an F1-score of 95.30%, but significantly increased test time and parameter size. Variant 4 showed similar accuracy but suffered from a further reduction in recall and an increase in test time, indicating lower efficiency of the spatial encoding in the CA structure for this task.

Overall, the complete AugEffNet model achieved the best performance across all metrics, with an accuracy of 96.60%, an F1-score of 96.28%, and the highest recall and precision. This demonstrates that each component—especially the MS-GCDA module, the LCH classifier, and the SE attention mechanism—works synergistically to enhance the model's superior recognition capability without introducing excessive computational overhead.

To provide a more intuitive assessment of the MS-GCDA module's influence on AugEffNet's disease recognition performance, the Gradient-weighted Class Activation Mapping(Grad-CAM) visualization technique was employed to generate feature activation maps across different model variants. As shown in Fig. 12, when the MS-GCDA module is ablated, the model predominantly attends to irrelevant background regions—such as leaf veins, illumination variations, and shaded areas—often resulting in inaccurate localization of disease symptoms. In contrast, the full AugEffNet model incorporating the MS-GCDA module focuses more

precisely on lesion regions, effectively suppressing redundant activations arising from background noise and enhancing the discriminability of disease-specific features. These observations demonstrate that integrating the MS-GCDA module directs the network's attention to pathologically salient areas, thereby strengthening both the robustness and interpretability of recognition outcomes.

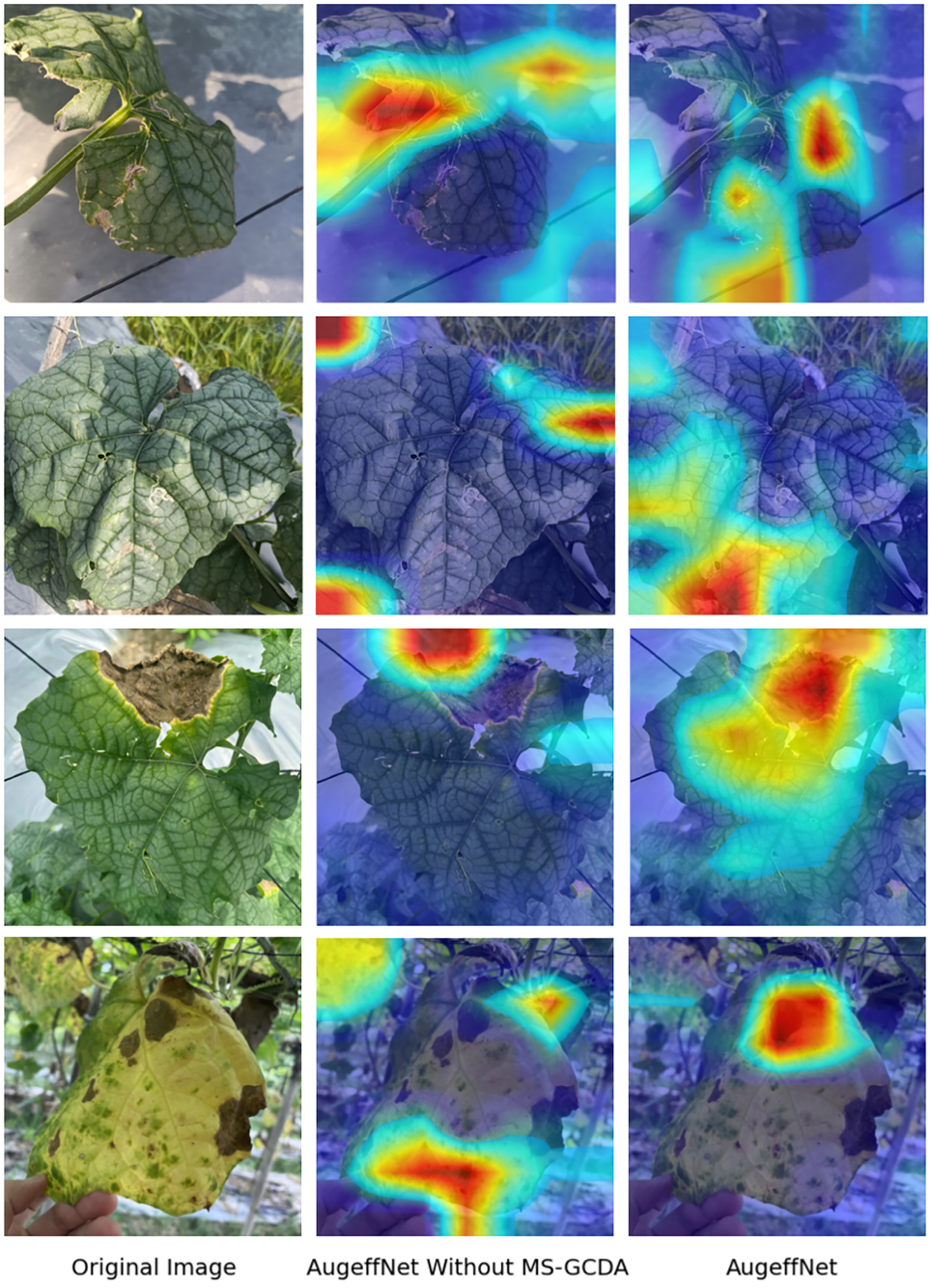

Figure 12: Feature visualization in various background.

3.5 *Effectiveness of Latent Feature Projection for Edge Deployment*

Table 7 evaluates the impact of the proposed Latent Feature Projection (LFP) module on both recognition performance and computational efficiency. By replacing the traditional high-dimensional dense classifier with a Learnable Linear Bottleneck, the framework achieves a superior trade-off between diagnostic accuracy and inference speed.

Table 7: Comparative analysis of LFP on model performance and efficiency.

Model	Accuracy (%)	F1-Score (%)	Recall (%)	Precision (%)	FLOPS (M)	Params (M)	Test time (ms)
AugEffNet	96.60	96.28	96.23	96.53	161.22	0.457	127.57
AugEffNet With LFP	97.27	97.05	97.09	97.09	122.88	3.28	6.68

The LFP module functions as a structural regularizer that adaptively projects the high-dimensional features from the MS-GCDA module into a task-specific 128-dimensional latent space. As shown in Table 7, this architectural refinement leads to a significant reduction in model complexity. Specifically, the total parameter count is reduced from 161.22 to 122.88M. Unlike static dimensionality reduction methods, the learnable nature of LFP ensures that critical diagnostic markers are preserved during feature compression, maintaining a competitive F1-score of 97.05 and an accuracy of 97.27.

The most substantial improvement is observed in real-world execution latency. The measured mean inference time is reduced from 127.57 to 6.68 ms, representing a performance gain of 95%. This acceleration is primarily attributed to the elimination of high-dimensional matrix multiplications in the classification head and the optimization of memory access patterns.

Our results indicate that the LFP-based classification strategy provides a more "deep-learning-consistent" solution compared to traditional non-learnable pipelines. By integrating feature projection directly into the backpropagation-optimized architecture, the proposed framework effectively mitigates the computational bottlenecks of resource-constrained UAV platforms while enhancing the discriminative stability of pest and disease recognition in complex field environments.

3.6 *Robustness Evaluation under Environmental Disturbances*

To assess the robustness of AugEffNet in practical UAV-based agricultural monitoring scenarios, we conducted controlled stress tests under representative environmental perturbations. Specifically, we benchmarked the proposed model against two competitive lightweight architectures, EdgeNeXt and MobileViTv2, under three categories of disturbances: Gaussian noise (simulating sensor degradation), Gaussian blur (capturing motion-induced artifacts), and scale variation (reflecting altitude fluctuations). The quantitative results are summarized in Table 8.

3.6.1 Resilience to Sensor Noise and Motion-Induced Degradation

As reported in Table 8, AugEffNet achieves the highest performance under clean conditions, with an accuracy of 97.27%, compared to 96.25% for EdgeNeXt and 95.90% for MobileViTv2. Under mild sensor noise (σ = 0.01), AugEffNet maintains a high accuracy of 97.07%, indicating strong robustness to low-level perturbations.

Table 8: Robustness evaluation under different disturbance types and settings. ΔAccuracy denotes the performance drop relative to the clean setting of each model.

Model	Disturbance Type	Setting	Accuracy (%)	Δ Acc	Precision (%)	Recall (%)	F1-Score (%)
EdgeNeXt	Clean	Original	96.25	0.00	95.87	95.76	96.01
	Gaussian Noise	σ = 0.01	95.56	−0.69	95.42	94.98	95.07
		σ = 0.05	91.30	−4.95	91.47	90.65	90.64
		σ = 0.10	73.38	−22.87	76.45	72.57	71.93
	Gaussian Blur	3 × 3	94.88	−1.37	94.50	94.32	94.30
		5 × 5	92.32	−3.93	92.16	91.64	91.72
		7 × 7	87.05	−9.20	88.50	86.66	87.06
	Scale Variation	0.5×	93.34	−2.91	93.05	92.71	92.76
		1.5×	95.73	−0.52	95.43	95.50	95.41
		2.0×	95.73	−0.52	95.35	95.22	95.23
MobileViTv2	Clean	Original	95.90	0.00	95.53	95.50	95.45
	Gaussian Noise	σ = 0.01	95.56	−0.34	95.42	94.98	95.07
		σ = 0.05	80.38	−15.52	83.11	78.20	78.35
		σ = 0.10	55.46	−40.44	72.89	51.15	46.46
	Gaussian Blur	3 × 3	93.17	−2.73	92.90	92.85	92.77
		5 × 5	88.40	−7.50	88.71	88.10	87.95
		7 × 7	79.18	−16.72	82.31	79.16	78.65
	Scale Variation	0.5×	89.76	−6.14	89.84	89.76	89.49
		1.5×	95.73	−0.17	95.35	95.50	95.41
		2.0×	95.73	−0.17	95.35	95.50	95.41
AugEffNet with LFP	Clean	Original	97.27	0.00	97.09	97.09	97.05
	Gaussian Noise	σ = 0.01	97.07	−0.20	96.75	96.83	96.91
		σ = 0.05	83.55	−13.72	79.42	81.21	80.20
		σ = 0.10	72.89	−24.38	73.01	71.70	70.13
	Gaussian Blur	3 × 3	94.37	−2.90	94.11	93.92	93.89
		5 × 5	90.78	−6.49	91.20	90.26	90.30
		7 × 7	89.01	−8.26	88.61	85.41	85.63
	Scale Variation	0.5×	91.30	−5.97	91.83	90.72	90.84
		1.5×	97.10	−0.17	96.82	96.95	96.86
		2.0×	97.02	−0.25	96.75	96.88	96.81

As noise intensity increases, all models exhibit performance degradation. At σ = 0.05, AugEffNet decreases to 83.55%, lower than EdgeNeXt (91.30%) but comparable to MobileViTv2 (80.38%). This sensitivity is likely due to the ultra-compact parameterization of AugEffNet, which limits redundancy under severe corruption. At σ = 0.10, AugEffNet (72.89%) performs comparably to EdgeNeXt (73.38%) and significantly outperforms MobileViTv2 (55.46%).

Under motion blur, AugEffNet demonstrates strong resilience. Although accuracy decreases with increasing blur kernel size, it consistently outperforms MobileViTv2 and remains close to EdgeNeXt. Notably, under severe blur (7×7), AugEffNet achieves 89.01%, exceeding MobileViTv2 by a substantial margin, suggesting that the MS-GCDA module effectively prioritizes structural semantic information over high-frequency details.

3.6.2 Scale Invariance and Robustness to Altitude Variations

Scale variation is a fundamental challenge in UAV-based recognition due to dynamic flight altitudes. The results demonstrate that AugEffNet exhibits strong scale invariance. When the input is upscaled (1.5× and 2.0×), the model maintains near-baseline performance, achieving 97.10% and 97.02%, respectively.

In contrast, EdgeNeXt and MobileViTv2 show limited improvement under scale enlargement, with accuracies remaining around 95.73%. Under downscaling (0.5×), AugEffNet achieves 91.30%, outperforming MobileViTv2 (89.76%) but slightly below EdgeNeXt (93.34%). These results indicate that while AugEffNet generalizes effectively across varying scales, extreme resolution reduction still affects fine-grained feature representation.

Overall, the results confirm that the proposed multi-scale grouped dual-attention mechanism enables adaptive receptive field calibration, ensuring stable performance across diverse spatial transformations and enhancing robustness in real-world UAV deployment scenarios.

3.7 Model Deployment

To support practical deployment in precision agriculture, the proposed framework is integrated into a UAV-assisted architecture for real-time pest and disease monitoring in loofah cultivation. As illustrated in Fig. 13, the overall system comprises periodic UAV patrols, video stream acquisition, edge-side analysis, and real-time visualization of recognition results, forming a closed-loop monitoring pipeline.

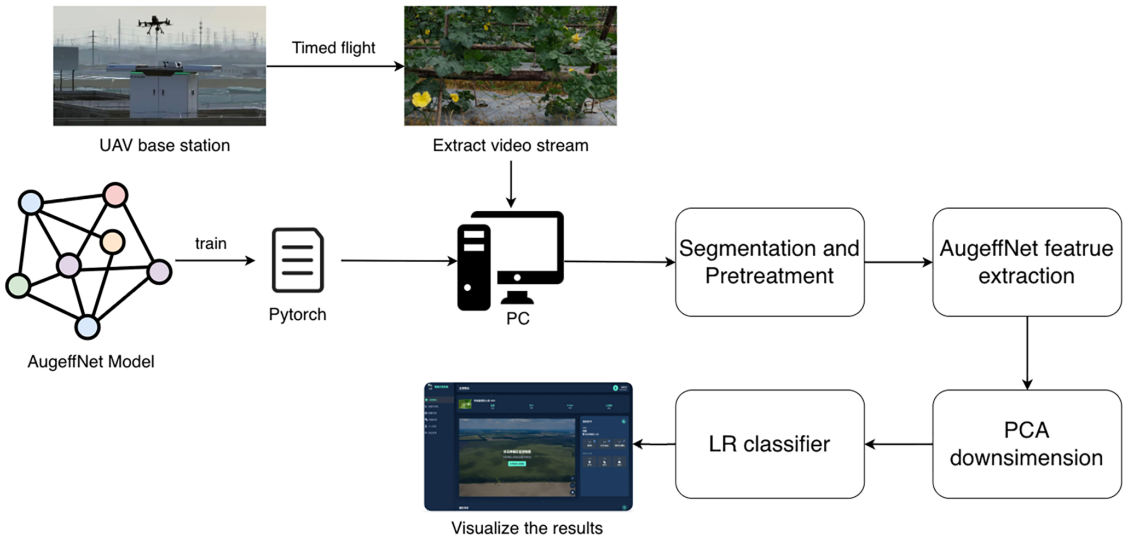

Figure 13: UAV edge deployment process.

To construct a readily deployable real-time monitoring system, this study implemented a complete UAV-ground station processing pipeline based on consumer-grade hardware platforms. The system employs a DJI

Mini 4 Pro UAV for field data acquisition, while real-time video stream analysis and disease identification are performed on a ground station laptop.

In the deployment scenario, a UAV equipped with an RGB camera conducts scheduled low-altitude flights over the target farmland along predefined routes. During each flight mission, the UAV captures continuous video streams of the loofah field for approximately 1 min and transmits the data wirelessly to a ground-based edge computing unit or a nearby workstation for subsequent analysis. Compared with satellite- or high-altitude-based remote sensing approaches, UAV-based close-range observation offers substantially higher spatial resolution and greater operational flexibility, making it particularly well suited for capturing fine-grained disease symptoms and small lesion regions on crop leaves.

Upon receiving the video stream, representative keyframes are extracted and forwarded to the preprocessing module. To mitigate the effects of complex, cluttered backgrounds commonly encountered in natural field environments, a leaf segmentation and pretreatment stage is first applied. This process isolates candidate leaf regions from soil, weeds, and supporting structures, thereby enabling the subsequent recognition model to focus on disease-relevant visual information. Fig. 14 illustrates representative leaf segmentation results obtained from UAV video frames.

Figure 14: Leaf segmentation in video stream.

In performance verification, the system processed a video sequence captured in a real loofah field. Test results show that the AugEffNet model maintains high classification accuracy (consistent with the experimental results in Section 3.1) while meeting real-time single-frame processing requirements. The deployed model effectively distinguishes healthy leaves from various pest and disease symptoms, demonstrating robustness even in complex scenarios with varying light levels and partial shading. This deployment verification demonstrates that the AugEffNet model can be integrated into edge computing devices without complex quantization or hardware-specific optimizations, achieving efficient real-time pest and disease identification and laying a technical foundation for its practical application in smart agriculture systems.

4 Discussions

Intelligent pest and disease recognition is a core component of smart agriculture and precision crop management. With the rapid development of remote sensing technologies, both satellite-based systems for regional-scale monitoring and unmanned aerial vehicle (UAV) platforms for field-level observation have been widely explored. Compared with satellite remote sensing, UAV-based monitoring provides higher spatial resolution and greater operational flexibility, making it particularly suitable for fine-grained assessment of crop health conditions [31].

Despite these advances, reliable disease recognition in real-world agricultural environments remains challenging [32]. Complex backgrounds, large variations in symptom scale, and subtle visual differences between healthy and diseased tissues place high demands on feature representation capability while simultaneously constraining computational efficiency. Within this context, the present study shifts the perspective from traditional modular engineering to a scale-robust cross-scale representation learning paradigm. Rather than pursuing increasingly complex network structures, the proposed AugEffNet framework emphasizes the synergy between scale-adaptive structural sampling and semantic importance rectification.

The core strength of the MS-GCDA module lies in its sequential "detect-then-rectify" strategy, which explicitly addresses the cross-scale feature misalignment inherent in UAV imaging. Our robustness evaluations confirm that the multi-scale spatial attention branch, utilizing parallel mapping functions (3×3 and 5×5 kernels), establishes exceptional scale invariance. The model maintains an accuracy exceeding 97% even under 1.5× and 2.0× scale variations, significantly surpassing the stability of larger architectures like EdgeNeXt and MobileViTv2. This suggests that for many foliar diseases, discriminative visual cues—such as lesions and discoloration patterns—are locally defined and can be effectively preserved across observation altitudes through principled architectural choices.

From a deployment perspective, the integration of Latent Feature Projection (LFP) and a learnable linear bottleneck further enhances the practicality of the proposed approach. By distilling high-dimensional MS-GCDA features into a compact 128-dimensional latent space, the framework achieves a "structural sparsity" that substantially lowers memory access and data movement overhead. Such system-level efficiency gains are particularly relevant for prolonged UAV missions, where power consumption and onboard computational capacity directly affect operational feasibility. This hardware-aware optimization aligns with recent trends in aerial vision research, such as the use of adaptive convolutions to handle orientation and scale instability [6].

In light of recent advancements in Vision Transformers, our comparison with SOTA models like TinyViT and MobileViTv2-1.0 reveals that while Transformer-based architectures offer strong global context modeling, their quadratic computational complexity remains a bottleneck for real-time UAV edge deployment. Our findings demonstrate that a formalized, lightweight CNN-based attention mechanism can achieve a superior accuracy-efficiency trade-off, reaching 96.60% accuracy with only 0.457 M parameters. However, we acknowledge that exploring hybrid Transformer-CNN architectures represents a promising future direction to further combine local efficiency with global relational modeling.

Nevertheless, several limitations warrant discussion. First, the absence of large-scale, fully annotated UAV disease datasets constrains direct validation under diverse aerial imaging conditions. Second, while the framework demonstrates robust transferability, the generalization to multi-crop and multi-region data is an ongoing challenge that involves complex cross-regional calibration. Furthermore, the present study focuses exclusively on RGB imagery and does not exploit multispectral or temporal information, which may further improve disease discrimination [31]. Future work will therefore explore the integration of multi-temporal observations and multispectral cues within this scale-robust framework to enable more comprehensive characterization of disease progression in dynamic farming environments.

In conclusion, this study demonstrates that scale-aware representation learning, combined with deployment-oriented optimization, constitutes a viable pathway for translating deep learning models into practical UAV-based agricultural remote sensing systems. By explicitly addressing scale variability, background complexity, and edge constraints, the proposed approach provides both methodological insights and practical guidance for the development of deployable intelligent plant protection systems.

5 Conclusion

This study introduces a scale-robust cross-scale representation learning paradigm designed to address target scale variability and background clutter in UAV-based agricultural monitoring. By implementing a systematic "detect-then-rectify" strategy, the proposed AugEffNet achieves an optimal balance between diagnostic precision and edge-computing efficiency. This approach provides a resilient solution for feature extraction under the dynamic environmental conditions characteristic of aerial remote sensing, effectively bridging the gap between deep learning theory and practical agricultural application.

The core MS-GCDA module effectively captures multi-scale spatial contexts and rectifies semantic importance through grouped channel attention, ensuring feature stability across significant altitude fluctuations. Extensive robustness evaluations demonstrate that the framework maintains high recognition accuracy under various scale variations and environmental disturbances. Furthermore, the Latent Feature Projection (LFP) strategy optimizes the model's structural footprint for real-time UAV deployment, achieving significant reductions in inference latency and memory overhead while preserving the essential representational power of deep features.

In summary, AugEffNet provides a high-performance, ultra-lightweight architecture that outperforms recent state-of-the-art methods in terms of accuracy-efficiency trade-offs. While this work establishes a rigorous foundation for autonomous plant protection, future research will focus on extending this framework to multi-crop and multi-region datasets to validate its cross-regional stability. Additionally, exploring hybrid architectures remains a priority to further enhance global modeling capabilities, ultimately offering a scalable and deployable solution for intelligent crop management in precision agriculture.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Conceptualization: Dingju Zhu, Shihua Mao, Jinchen Wu, Depeng Kong, Kaileung Yung and Andrew W. H. Ip; methodology: Kemeng Zhu, Dingju Zhu, Jinchen Wu, Depeng Kong, Kaileung Yung and Andrew W. H. Ip; formal analysis: Kemeng Zhu; investigation: Shihua Mao; data curation: Kemeng Zhu and Shihua Mao; software: Kemeng Zhu; visualization: Shihua Mao; writing—original draft preparation: Kemeng Zhu; writing—review and editing: Kemeng Zhu, Dingju Zhu and Shihua Mao; supervision: Dingju Zhu. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets supporting the conclusions of this article are publicly available. The AppleLeaf9 dataset is hosted in a GitHub repository at: https://github.com/JasonYangCode/AppleLeaf9. The self-collected Loofah dataset is published in Mendeley Data at: https://data.mendeley.com/datasets/wyhm534ptg/1.

Ethics Approval: Not applicable.

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

References

1. Shakhatreh H, Sawalmeh AH, Al-Fuqaha A, Dou Z, Almaita E, Khalil I, et al. Unmanned Aerial Vehicles (UAVs): a survey on civil applications and key research challenges. IEEE Access. 2019;7:48572–634. doi:10.1109/ACCESS.2019.2909530.
2. Mohammad-Razdari A, Rousseau D, Bakhshipour A, Taylor S, Poveda J, Kiani H. Recent advances in E-monitoring of plant diseases. Biosens Bioelectron. 2022;201(1):113953. doi:10.1016/j.bios.2021.113953.
3. Bao W, Yang X, Liang D, Hu G, Yang X. Lightweight convolutional neural network model for field wheat ear disease identification. Comput Electron Agric. 2021;189(4):106367. doi:10.1016/j.compag.2021.106367.
4. Mohanty SP, Hughes DP, Salathé M. Using deep learning for image-based plant disease detection. Front Plant Sci. 2016;7:01419. doi:10.3389/fpls.2016.01419.
5. Lafta M. Path scheduling and target trajectory optimization in UAVs based on dragonfly and firefly algorithm. Adv Eng Intell Syst. 2022;10:001. doi:10.22034/aeis.2022.356205.1037.
6. Umirzakova S, Muksimova S, Mahliyo Olimjon Qizi A, Cho YI. Lightweight transformer with adaptive rotational convolutions for aerial object detection. Appl Sci. 2025;15(9):5212. doi:10.3390/app15095212.
7. Fan Z, Lu D, Liu M, Liu Z, Dong Q, Zou H, et al. YOLO-PDGT: a lightweight and efficient algorithm for unripe pomegranate detection and counting. Measurement. 2025;254(4):117852. doi:10.1016/j.measurement.2025.117852.
8. Nazari K, Ebadi MJ, Berahmand K. Diagnosis of Alternaria disease and leafminer pest on tomato leaves using image processing techniques. J Sci Food Agric. 2022;102(15):6907–20. doi:10.1002/jsfa.12052.
9. Yang Q, Duan S, Wang L. Efficient identification of apple leaf diseases in the wild using convolutional neural networks. Agronomy. 2022;12(11):2784. doi:10.3390/agronomy12112784.
10. Zhang J, Liu Z, Yu K. MSFNet-CPD: multi-scale cross-modal fusion network for crop pest detection. arXiv:2505.02441. 2025.
11. Dong P, Li K, Wang M, Li F, Guo W, Si H. Maize leaf compound disease recognition based on attention mechanism. Agriculture. 2024;14(1):74. doi:10.3390/agriculture14010074.
12. Mittal P. A comprehensive survey of deep learning–based lightweight object detection models for edge devices. Artif Intell Rev. 2024;57(9):1–47. doi:10.1007/s10462-024-10877-1.
13. Guan H, Fu C, Zhang G, Li K, Wang P, Zhu Z. A lightweight model for efficient identification of plant diseases and pests based on deep learning. Front Plant Sci. 2023;14:1227011. doi:10.3389/fpls.2023.1227011.
14. Yang X, Wang H, Zhou Q, Lu L, Zhang L, Sun C, et al. A lightweight and efficient plant disease detection method integrating knowledge distillation and dual-scale weighted convolutions. Algorithms. 2025;18(7):433. doi:10.3390/a18070433.
15. Feng J, Chao X. Apple tree leaf disease segmentation dataset. Beijing, China: Science Data Bank; 2022. doi:10.11922/sciencedb.01627.
16. Thapa R, Zhang K, Snavely N, Belongie S, Khan A. The plant pathology challenge 2020 data set to classify foliar disease of apples. Appl Plant Sci. 2020;8(9):e11390. doi:10.1002/aps3.11390.
17. Thapa R, Zhang K, Snavely N, Belongie S, Khan A. Plant pathology 2021—FGVC8. Kaggle. 2021 [cited 2026 Jan 1]. Available from: https://kaggle.com/competitions/plant-pathology-2021-fgvc8.
18. Li X, Wang W, Hu X, Yang J. Selective kernel networks. arXiv:1903.06586. 2019.
19. Gao SH, Cheng MM, Zhao K, Zhang XY, Yang MH, Torr P. Res2Net: a new multi-scale backbone architecture. IEEE Trans Pattern Anal Mach Intell. 2021;43(2):652–62. doi:10.1109/tpami.2019.2938758.
20. Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-excitation networks. arXiv:1709.01507. 2019.
21. Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. arXiv:1905.11946. 2020.
22. Ma N, Zhang X, Zheng HT, Sun J. ShuffleNet V2: practical guidelines for efficient CNN architecture design. arXiv:1807.11164. 2018.
23. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. arXiv:1608.06993. 2018.
24. Howard A, Sandler M, Chu G, Chen LC, Chen B, Tan M, et al. Searching for MobileNetV3. arXiv:1905.02244. 2019.
25. Han K, Wang Y, Tian Q, Guo J, Xu C, Xu C. GhostNet: more features from cheap operations. arXiv:1911.11907. 2020.

26. Chen J, Hong Kao S, He H, Zhuo W, Wen S, Lee CH, et al. Run, don't walk: chasing higher flops for faster neural networks. arXiv:2303.03667. 2023.
27. Luong HH. Improving potato diseases classification based on custom ConvNeXtSmall and combine with the explanation model. Int J Adv Comput Sci Appl. 2024;15(4):01504121. doi:10.14569/IJACSA.2024.01504121.
28. Maaz M, Shaker A, Cholakkal H, Khan S, Zamir SW, Anwer RM, et al. EdgeNeXt: efficiently amalgamated CNN-transformer architecture for mobile vision applications. arXiv:2206.10589. 2022.
29. Yaseen M. What is YOLOv8: an in-depth exploration of the internal features of the next-generation object detector. arXiv:2408.15857. 2024.
30. Wu K, Zhang J, Peng H, Liu M, Xiao B, Fu J, et al. TinyViT: fast pretraining distillation for small vision transformers. arXiv:2207.10666. 2022.
31. Petso T, Jamisola R. A review on deep learning on UAV monitoring systems for agricultural applications. Artif Intell Robot Auton Syst Appl. 2023;1093(5):335–68. doi:10.1007/978-3-031-28715-2_11.
32. Joshi H. Edge-AI for agriculture: lightweight vision models for disease detection in resource-limited settings. arXiv:2412.18635. 2024.