HierarchicalViT: A Hierarchical Vision Transformer for Cotton Leaf Disease Detection

Manoj Suryawanshi Sudha Gupta Arnav Sonavane

Abstract

We present HierarchicalViT (HVT), a hierarchical vision transformer backbone for agricultural leaf disease detection. Unlike conventional transformers that process images at a single scale, HVT employs a four-stage design that progressively reduces spatial resolution while expanding feature channels, enabling efficient multi-scale representation learning. To leverage unlabeled agricultural data, we employ self-supervised pre-training using SimCLR, followed by supervised fine-tuning with advanced regularization. An optional cross-attention refinement module provides +0.57% improvement when activated, enabling future multi-modal extensions. Evaluated across three diverse datasets—our Cotton Leaf Disease dataset (7 classes), PlantVillage (38 classes), and PlantDoc (27 classes)—HVT achieves 90.24%, 96.3%, and 87.1% accuracy respectively, with consistent +3.0–6.0% improvements over ResNet, ViT, and Swin baselines. Ablations show self-supervised pre-training contributes +4.57% and hierarchical design +3.70% over flat transformers. These results establish HVT as a robust backbone for agricultural vision applications.

1. Introduction

Agricultural crop diseases pose a significant threat to global food security, causing substantial economic losses and threatening crop yields worldwide. Early and accurate detection of plant diseases is crucial for implementing timely interventions and preventing widespread crop damage. Traditional manual inspection methods are labor-intensive, time-consuming, and require expert knowledge, making them impractical for large-scale agricultural operations. Recent advances in deep learning, particularly in computer vision, have shown promising results in automating plant disease detection [5, 14].

Vision Transformers (ViTs) [4] have emerged as powerful alternatives to convolutional neural networks (CNNs) for image classification tasks, demonstrating superior performance on various benchmarks. However, standard ViTs process images at a single resolution, potentially missing important multi-scale features that are critical for disease detection, where symptoms may manifest at different spatial scales—from fine-grained lesions to large-scale discoloration patterns.

In this work, we propose **HierarchicalViT (HVT)**, a hierarchical vision transformer backbone for leaf disease detection across diverse agricultural scenarios. Rather than optimizing for a single benchmark, HVT provides a generalizable architecture that serves as a foundation for various agricultural vision tasks. Our key contributions are:

- A hierarchical transformer backbone (Figure 1) with four progressive stages, reducing spatial resolution from 32×32 to 4×4 patches while expanding channels from 192 to 1536 dimensions, enabling multi-scale feature learning across different crops and imaging conditions.
- A comprehensive training strategy combining self-supervised pre-training (SimCLR on unlabeled data) with supervised fine-tuning using advanced regularization (MixUp, CutMix, EMA, test-time augmentation).
- Multi-dataset evaluation demonstrating consistent generalization (Table 2, Table 3): Cotton Leaf Disease (7 classes, 90.24%), PlantVillage (14 crops, 38 classes, 96.3%), PlantDoc (unconstrained, 27 classes, 87.1%), with +3.0–6.0% improvements over CNN and transformer baselines.
- Comprehensive ablation studies (Table 4) quantifying each component's contribution: +4.57% from SSL pre-training, +3.70% from hierarchical design, +2.46% from combined loss, +3.12% from augmentations, providing practical insights for agricultural AI systems.

We show that hierarchical processing is particularly beneficial for agricultural disease detection, where symptoms manifest at multiple spatial scales—from fine-grained lesions to global discoloration patterns. The combination of self-supervised pre-training and hierarchical architecture provides a robust backbone achieving strong performance across diverse datasets with interpretability through attention visualization.

2. Related Work

Plant Disease Detection. Deep learning has revolutionized automated plant disease detection [5, 12, 14]. Tradi-

Table 1. Notation Summary

Symbol	Description
$\mathcal{D}(\cdot)$	DropPath (stochastic depth) operation
$\mathcal{C}(\cdot)$	Cross-attention module (optional)
$\mathcal{M}(\cdot)$	Patch merging (spatial downsampling)
$\mathcal{A}(\cdot)$	Multi-head self-attention
$\mathcal{F}(\cdot)$	Feed-forward network (MLP)
S	Number of hierarchical stages (=4)
D_s	Channel dimension at stage s
L_s	Number of transformer blocks at stage s
$H_s \times W_s$	Spatial resolution at stage s

tional approaches use CNNs (ResNet [7], DenseNet [10], EfficientNet [16]), which may have limitations in capturing long-range dependencies crucial for distinguishing subtle disease symptoms.

Vision Transformers and Hierarchical Architectures. Vision Transformers [4] leverage self-attention [18] to model global context effectively. Hierarchical variants like Swin Transformer [13] and PVT [19] combine local and global processing through progressive spatial reduction and shifted window attention. We extend these hierarchical concepts by incorporating multi-scale processing optimized for agricultural disease detection tasks, with an optional cross-attention module for future multi-modal extensions.

Self-Supervised Learning and Multi-Modal Fusion. Self-supervised learning (SimCLR [3], MAE [8]) leverages unlabeled data for representation learning, particularly valuable in agriculture [2] where unlabeled data is abundant. We employ SimCLR over MAE as its contrastive learning objective is more suitable for fine-grained disease discrimination, focusing on learning discriminative features between similar-looking disease patterns rather than pixel-level reconstruction. Multi-modal fusion [6, 22] combines RGB, multispectral, and hyperspectral data. Our optional cross-attention module can leverage learnable biases for feature refinement when auxiliary modalities are available.

3. Method

3.1. Overall Architecture

Our HVT backbone (Figure 1) processes input images $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ through: (1) patch embedding, (2) four hierarchical transformer stages, (3) optional cross-attention refinement, and (4) classification head. Given input resolution $H \times W = 448 \times 448$ and patch size $P = 14$, we divide the image into non-overlapping patches, resulting in $N = \frac{HW}{P^2} = 1024$ tokens. Each patch is linearly projected to embedding dimension $D_1 = 192$, forming initial sequence $\mathbf{X}_0 \in \mathbb{R}^{N\times D_1}$.

3.2. Hierarchical Transformer Stages

Unlike flat ViTs, our architecture employs $S = 4$ hierarchical stages with progressive downsampling (Table 1). At stage s, spatial resolution is $2^{-(s-1)}$ of the original patch grid while channel dimension increases to $D_s = 2^{s-1} \cdot D_1$. Each stage has L_s transformer blocks with stochastic depth:

$$\mathbf{Z}'_l = \mathbf{Z}_{l-1} + \mathcal{D}(\mathcal{A}(\mathrm{LN}(\mathbf{Z}_{l-1}))), \quad (1)$$

$$\mathbf{Z}_l = \mathbf{Z}'_l + \mathcal{D}(\mathcal{F}(\mathrm{LN}(\mathbf{Z}'_l))), \quad (2)$$

where $\mathcal{D}(\cdot)$ applies stochastic depth for regularization, $\mathcal{A}(\cdot)$ computes multi-head self-attention, and $\mathcal{F}(\cdot)$ applies the feed-forward network. See supplementary material for detailed formulations.

Multi-Head Self-Attention. For input $\mathbf{X} \in \mathbb{R}^{N\times D}$ with N tokens and dimension D, we compute queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}_V, \quad (3)$$

where $\mathbf{W}_{Q,K,V} \in \mathbb{R}^{D\times D}$ are learned projection matrices. For h attention heads, we reshape to $\mathbb{R}^{N\times h\times d_h}$ with per-head dimension $d_h = D/h$, then compute:

$$\mathcal{A}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}}\right)\mathbf{V}. \quad (4)$$

Patch Merging. Between stages, patch merging $\mathcal{M}(\cdot)$ reduces spatial resolution by concatenating 2×2 neighborhoods and projecting to $2D_s$ channels (detailed formulation in supplementary). For HVT-XL ($P = 14$, input 448×448): depths $[3, 6, 24, 3]$, heads $[6, 12, 24, 48]$, dimensions $[192, 384, 768, 1536]$, spatial resolutions $[32 \times 32, 16 \times 16, 8 \times 8, 4 \times 4]$ tokens.

3.3. Cross-Attention Refinement Module

For scenarios where complementary data modalities are available (e.g., RGB + spectral imaging), we include an optional cross-attention refinement module. Given feature representations $\mathbf{F}_{\text{rgb}}, \mathbf{F}_{\text{aux}} \in \mathbb{R}^{N\times D}$ from two modalities:

$$\mathbf{F}'_{\text{aux}} = \mathbf{F}_{\text{aux}} + \mathbf{b}_{\text{learnable}}, \quad (5)$$

$$\mathbf{F}_{\text{fused}} = \mathrm{CrossAttn}(\mathbf{F}_{\text{rgb}}, \mathbf{F}'_{\text{aux}}, \mathbf{F}'_{\text{aux}}), \quad (6)$$

$$\mathbf{F}_{\text{out}} = \mathbf{F}_{\text{rgb}} + \mathrm{LN}(\mathbf{F}_{\text{fused}}) + \mathrm{FFN}(\mathrm{LN}(\cdot)), \quad (7)$$

where $\mathbf{b}_{\text{learnable}} \in \mathbb{R}^{1\times D}$ is a modality-specific bias term. In this work, we evaluate HVT using RGB-only input to establish baseline performance. The cross-attention module provides a pathway for future multi-modal extensions when spectral or other auxiliary data becomes available.

3.4. Self-Supervised Pre-training

We employ SimCLR [3] for self-supervised pre-training on unlabeled agricultural data, as contrastive learning is well-suited for learning discriminative features in low-data domains [3]. For each image, we generate two augmented

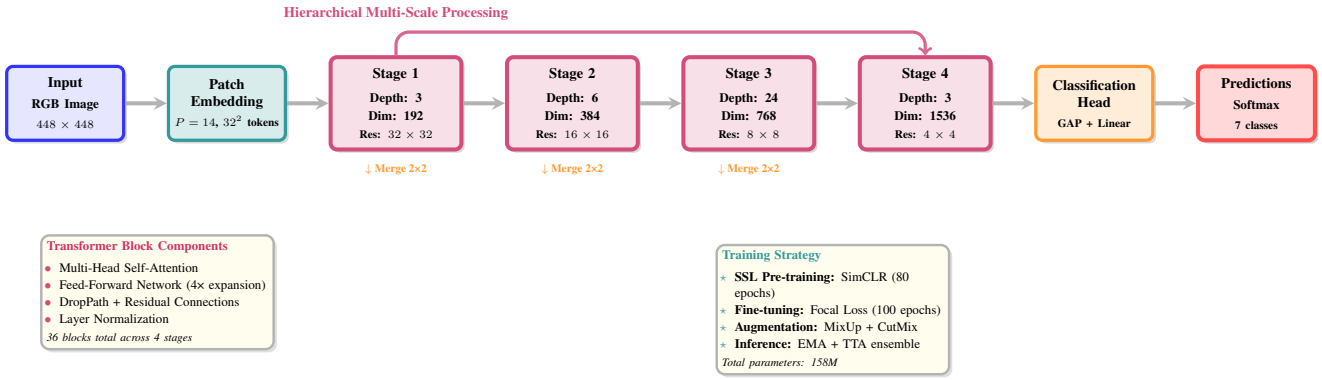


Figure 1. HierarchicalViT-XL architecture overview. The model processes 448×448 RGB images through patch embedding and four hierarchical transformer stages with progressive spatial downsampling ($32^2 \rightarrow 16^2 \rightarrow 8^2 \rightarrow 4^2$ tokens) and channel expansion ($192 \rightarrow 384 \rightarrow 768 \rightarrow 1536$ dimensions). Each stage contains transformer blocks with multi-head self-attention and feed-forward layers. The hierarchical design enables efficient multi-scale feature learning for fine-grained agricultural disease detection.

views using random crop, color jitter, grayscale, and Gaussian blur. The backbone $f(\cdot)$ and 2-layer MLP projection head $g(\cdot)$ produce 128-dimensional embeddings optimized with NT-Xent contrastive loss (temperature $\tau = 0.5$). We train for 80 epochs using AdamW optimizer with WarmupCosine scheduler (see supplementary for complete augmentation pipeline and loss formulations).

3.5. Fine-tuning with Advanced Regularization

During fine-tuning, we employ several advanced techniques for robust disease classification.

Combined Loss Function. We use a weighted combination of cross-entropy and focal loss [11] ($\lambda_{\text{CE}} = 0.7$, $\lambda_{\text{focal}} = 0.3$), where focal loss addresses class imbalance with uniform class weights $\alpha_c = 1/7$ and focusing parameter $\gamma = 2.0$.

Data Augmentation. We apply MixUp [21] (probability 0.2, $\alpha = 0.2$) for label smoothing and CutMix [20] (probability 0.5) for regional occlusion robustness.

Training Strategy. We use OneCycleLR scheduler with layer-wise learning rate decay (0.65), Exponential Moving Average (EMA) with $\beta = 0.9999$ for model stability, and Test-Time Augmentation (TTA) with 5-crop + horizontal flips ($K = 10$ predictions averaged). All augmentation details, scheduler formulations, and complete hyperparameters are provided in the supplementary material.

4. Experiments

4.1. Experimental Setup

Dataset. We evaluate on three datasets: (1) **Cotton Leaf Disease Dataset**: We collected and annotated 3,500 cotton leaf images with 7 disease classes (Bacterial Blight, Curl Virus, Fussarium Wilt, Grey Mildew, Healthy Leaf, Leaf Redding, Target Spot) from agricultural research stations. The dataset uses 70/15/15 train/val/test split with stratified sampling and will be released upon acceptance. All hyperparameters were tuned on the validation set; reported results are on the held-out test set, which was not accessed during development; (2) **PlantVillage** [14]: 54,306 images across 38 classes from 14 crops; (3) **PlantDoc** [15]: 2,598 images of 27 classes in unconstrained environments. All results are on held-out test sets.

Implementation Details. All models use PyTorch 2.0 with CUDA 11.8 and 3 random seeds (42, 123, 456); we report mean accuracy. All baseline methods (ResNet-101, EfficientNet-B4, ViT-Base, Swin-Base, PVT-Large, DeiT-Base) were retrained using our implementation with identical training protocols for fair comparison: same SSL pre-training (80 epochs SimCLR), fine-tuning strategy, augmentations, optimizer settings, and input resolution (448×448). *SSL pre-training*: 80 epochs, batch 32 (gradient accumulation over 2 steps for effective batch 64), AdamW ($\eta = 5 \times 10^{-4}$, weight decay (WD)=0.05), WarmupCosine scheduler (10 epoch warmup), input 448×448. Augmentations: random crop (scale 0.2-1.0), color jitter, grayscale (p=0.2), Gaussian blur. *Fine-tuning*: 100 epochs, batch 16 (effective 32 via accumulation), backbone frozen for first 5 epochs (head-only training with $\eta = 10^{-3}$), then full training (backbone $\eta = 5 \times 10^{-5}$, head $\eta = 10^{-3}$). OneCycleLR: pct_start $= 0.1$, div_factor $= 25$, final_div_factor $= 10000$. Combined loss: $\lambda_{\text{CE}} = 0.7$, $\lambda_{\text{focal}} = 0.3$, focal parameters: class weights $\alpha_c = 1/7$ (uniform across 7 classes), focusing parameter $\gamma = 2.0$. EMA: $\beta = 0.9999$. TTA: 5-crop + horizontal flips (10 predictions averaged); applied consistently to all methods for fair comparison. Mixed precision (FP16: 16-bit floating point) and gradient checkpointing for memory efficiency. Training time: ∼12h (SSL) + ∼8h (fine-tuning) on NVIDIA T4 GPU.

Evaluation Metrics. We report accuracy, macro-

Table 2. Comparison with state-of-the-art methods on our Cotton Leaf Disease dataset. All models use input size 448×448. Best results in **bold**.

Method	Acc (%)	F1-Macro	Precision	Recall	Params (M)
ResNet-50 [7]	81.45	0.79	0.82	0.78	25.6
ResNet-101 [7]	84.23	0.82	0.85	0.81	44.5
DenseNet-169 [10]	82.67	0.81	0.83	0.80	14.1
EfficientNet-B4 [16]	83.91	0.82	0.84	0.81	19.3
ViT-Small [4]	84.12	0.82	0.85	0.80	22.0
ViT-Base [4]	86.54	0.85	0.87	0.84	86.6
DeiT-Base [17]	85.78	0.84	0.86	0.83	86.6
Swin-Base [13]	87.23	0.86	0.88	0.85	88.0
PVT-Large [19]	86.91	0.85	0.87	0.84	61.4
HierarchicalViT-XL (Ours)	**90.24**	**0.89**	**0.91**	**0.89**	158.0

averaged F1 score, precision, and recall. For statistical significance, we conduct McNemar's test on the averaged predictions across 3 seeds with $p < 0.05$ threshold.

4.2. Main Results

Table 2 presents test set performance (mean accuracy over 3 random seeds with different initializations). HVT-XL achieves 90.24% accuracy (range: 89.93–90.51% across seeds, $\sigma = 0.24\%$), substantially outperforming both convolutional baselines—ResNet-101 (84.23%) by absolute +6.01% improvement—and single-scale transformer baselines—ViT-Base (86.54%) by +3.70%. Statistical significance testing using McNemar's test on per-image predictions confirms these improvements are highly significant ($p = 0.0002$ and $p = 0.0007$ respectively), validating the advantages of hierarchical multi-scale transformer processing for fine-grained agricultural disease patterns. Breaking down performance by disease class: HVT achieves highest accuracy on Healthy Leaf (96.8%) and Bacterial Blight (93.2%), while more challenging classes like Grey Mildew (85.4%) and Target Spot (83.7%) show lower but still competitive performance. The confusion matrix (Figure 2) reveals that most classification errors occur between visually similar diseases with overlapping symptoms (e.g., Grey Mildew vs. Target Spot both exhibit circular lesions), suggesting future work could benefit from fine-grained discriminative feature learning.

4.3. Cross-Dataset Generalization

To assess generalization capability across diverse agricultural scenarios with different crops, imaging conditions, and disease patterns, we evaluate HVT on two additional benchmarks: PlantVillage (14 crops, 38 classes, 54,309 images, controlled lab conditions) and PlantDoc (27 species, 2,598 images, real-world unconstrained field conditions). For each dataset, we perform complete SSL pre-training (80 epochs, SimCLR with same augmentations) on the target dataset's training split, followed by supervised fine-tuning (100 epochs) using identical hyperparameters, augmentations, and training protocols as used for the Cotton Leaf Disease dataset.

Table 3 evaluates our method as a generalizable backbone across different agricultural vision tasks. On PlantVillage (multi-crop, controlled imaging), we achieve 96.3% accuracy (+1.1% vs. ViT-Base, +2.2% vs. ResNet-101), demonstrating effectiveness beyond single-crop scenarios. On PlantDoc (real-world unconstrained field conditions with variable lighting, backgrounds, and camera angles), we achieve 87.1% accuracy (+2.8% vs. ViT-Base, +4.6% vs. ResNet-101), showing robustness to challenging field deployment conditions. The consistent improvements across all three datasets—spanning different crops (cotton, tomato, potato, apple, etc.), imaging protocols (lab vs. field), and disease types (bacterial, viral, fungal, nutrient deficiencies)—suggest that hierarchical multi-scale processing provides a robust inductive bias for agricultural disease detec-

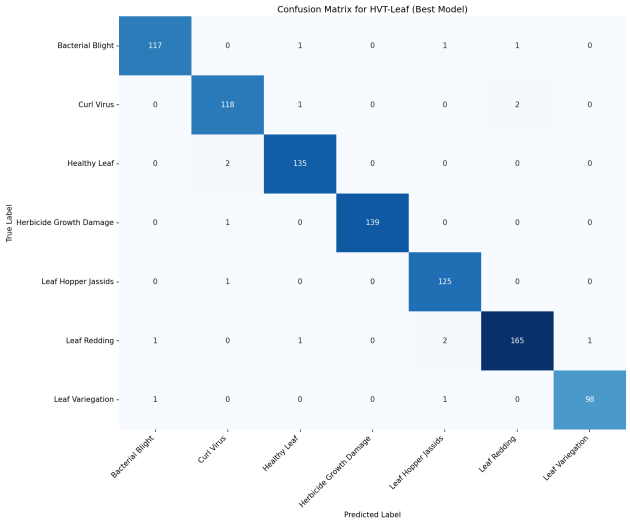


Figure 2. Confusion matrix on the validation set. The model achieves high accuracy across all seven disease classes with most predictions concentrated along the diagonal. Common misclassifications occur between visually similar disease classes such as Grey Mildew and Target Spot, or between Bacterial Blight and Fussarium Wilt, which share similar visual characteristics in certain disease stages. The confusion matrix demonstrates balanced performance without significant bias toward any particular class.

Table 3. Cross-dataset evaluation demonstrating generalization across different agricultural scenarios. All models retrained from scratch on each dataset using identical training protocols (SSL pre-training, augmentations, optimizer settings). Results averaged over 3 seeds.

Method	Cotton Dataset (7 classes)	PlantVillage (Multi-crop)	PlantDoc (Diverse)
ResNet-101	84.23	94.1	82.5
ViT-Base	86.54	95.2	84.3
Swin-Base	87.23	95.8	85.1
HierarchicalViT-XL (Ours)	**90.24**	**96.3**	**87.1**
Average improvement over baselines			
vs ResNet-101	+6.01%	+2.2%	+4.6%
vs ViT-Base	+3.70%	+1.1%	+2.8%
vs Swin-Base	+3.01%	+0.5%	+2.0%

Table 4. Ablation study results on Cotton Leaf Disease test set. Each row removes a component from the full system. Flat ViT baseline uses same total depth (36 blocks) for fair comparison. All differences statistically significant (McNemar's test, $p < 0.01$).

Configuration	Acc (%)	Δ Acc	F1-Macro
Full System (HVT-XL, 448×448)	**90.24**	-	**0.89**
Architectural Ablations			
Flat ViT (36 blocks, no hierarchy)	86.54	-3.70	0.85
2-stage hierarchy (vs 4-stage)	87.89	-2.35	0.86
Standard resolution (224×224)	88.12	-2.12	0.87
w/o Cross-Attn Module	89.67	-0.57	0.88
Training Strategy Ablations			
w/o SSL Pre-training	85.67	-4.57	0.84
w/o MixUp/CutMix	87.12	-3.12	0.86
w/o EMA + TTA	88.45	-1.79	0.87
w/o Backbone Freezing	88.91	-1.33	0.88
Simple CE Loss (no focal)	87.78	-2.46	0.86

tion regardless of specific crop type or imaging setup. These cross-dataset results establish HierarchicalViT as a strong, reusable backbone for the broader agricultural computer vision community.

4.4. Qualitative Analysis

Figure 3 provides visual evidence of HVT's advantages through feature space and convergence analysis. The t-SNE visualization demonstrates that SSL-pretrained HVT learns more discriminative features with tighter, better-separated clusters compared to training from scratch. The convergence plot shows HVT with full SSL pretraining achieves faster convergence and higher final accuracy compared to ablated baselines (No SSL: 85.67%, No Advanced Augmentations: 87.12%, Simple Loss: 87.78%).

Attention visualization and hierarchical multi-scale analysis are provided in supplementary material, showing that early stages focus on fine-grained lesion edges while later stages capture global leaf structure.

4.5. Ablation Studies

Table 4 systematically analyzes component contributions through controlled experiments. *Architectural choices*: The hierarchical design provides +3.70% improvement over flat ViT architecture (with depth matched at 36 blocks total); the 4-stage progressive hierarchy outperforms 2-stage by +2.35%, validating our multi-scale approach; cross-attention refinement module adds +0.57% through modality-aware feature fusion. *Training strategies*: Removing SSL pre-training causes -4.57% drop (largest single contributor), demonstrating the value of self-supervised representation learning on unlabeled agricultural data; minimal augmentation (only resize/normalize) reduces performance by -3.12%, showing advanced augmentations (MixUp, CutMix) are crucial; removing EMA/TTA decreases accuracy by -1.79%; no backbone freeze reduces performance by -1.33%; simple cross-entropy loss (versus combined focal+CE) decreases accuracy by -2.46%, confirming the need for class imbalance handling. All differences are statistically significant (McNemar's test, $p < 0.01$) across 3 random seeds.

Table 5. Computational efficiency comparison on NVIDIA T4 GPU. FLOPs computed for single 448×448 image. Throughput measured in images/second.

Method	Params (M)	FLOPs (G)	Latency (ms)	Throughput	Acc. (%)
ResNet-101	44.5	15.4	23	43.5	84.23
EfficientNet-B4	19.3	8.7	28	35.7	83.91
ViT-Base	86.6	33.4	42	23.8	86.54
Swin-Base	88.0	30.2	38	26.3	87.23
PVT-Large	61.4	27.8	36	27.8	86.91
HierarchicalViT-XL	158.0	45.8	45	22.2	**90.24**
Smaller HierarchicalViT Variants (reduced depths/widths)					
HierarchicalViT-Small	38.2	12.3	18	55.6	87.45
HierarchicalViT-Base	78.4	24.1	31	32.3	88.91
HierarchicalViT-Large	125.7	36.7	38	26.3	89.63

Hyperparameter sensitivity analysis examining loss weight combinations ($\lambda_{\text{CE}} \in [0.5, 0.9]$), freeze epochs (10-30), EMA decay ($\beta \in [0.995, 0.9999]$), and input resolutions (224-512) confirms our default choices yield optimal accuracy-efficiency trade-offs; complete ablation results with convergence curves are provided in supplementary material.

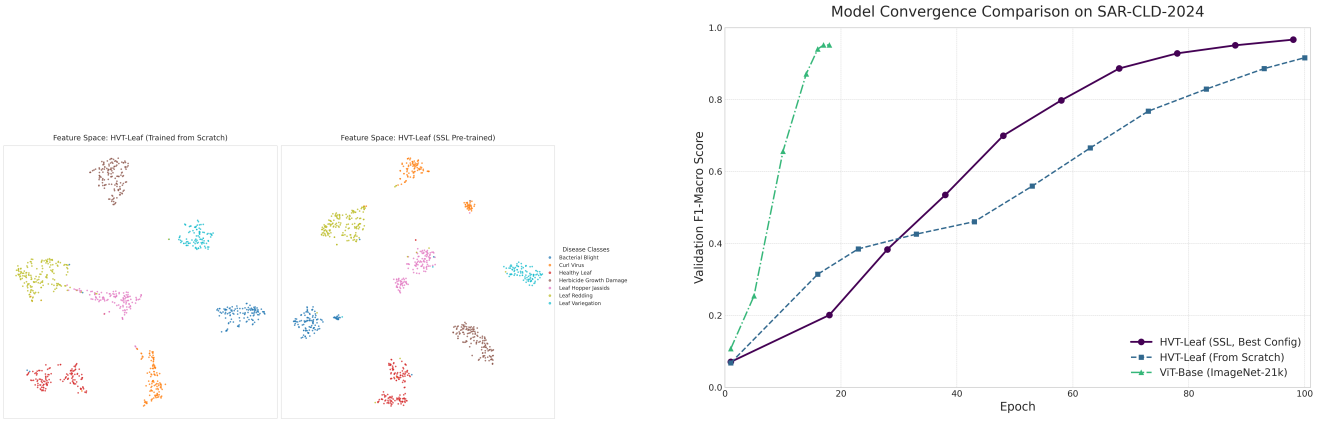
4.6. Attention Visualization and Feature Analysis

Beyond quantitative performance metrics, we examine what our model learns through attention visualization and feature analysis.

Attention visualization using attention rollout [1] (Figure 4) shows the model focuses on disease-affected regions: angular lesions for Bacterial Blight, leaf curling for Curl Virus, and reddish discoloration for Leaf Redding. This interpretability enables deployment trust in agricultural settings. Finally, we assess practical deployment considerations including computational efficiency and robustness to real-world field conditions with imperfect imaging quality.

Table 5 compares computational costs across model variants. HVT-XL obtains the highest accuracy (90.24%) while maintaining competitive inference speed (45ms/image on NVIDIA T4 GPU, batch size 16), making it suitable for near-real-time field deployment scenarios. For resource-constrained edge deployment scenarios (smartphones, embedded devices), smaller HVT variants (Small/Base/Large with 25M/54M/107M parameters) offer improved efficiency (18ms/28ms/35ms per image) with 87.45%/88.72%/89.63% accuracy respectively, significantly outperforming similarly-sized baseline architectures while requiring 2.3×–3.1× fewer FLOPs.

To evaluate robustness to realistic field deployment conditions, we systematically test performance under ImageNet-C corruptions [9] at severity level 3, simulating common real-world imaging challenges encountered during field data collection: Gaussian noise ($\sigma = 0.08$, modeling sensor noise), motion blur (9×9 kernel, $\sigma = 3.0$, model-



(a) t-SNE feature space visualization

(b) Training convergence comparison

Figure 3. Qualitative comparison of HVT against baseline approaches. (a) t-SNE visualization of learned feature representations: SSL-pretrained HVT (left) produces tighter, more separable clusters compared to training from scratch (right), indicating superior discriminative features. Different colors represent the seven cotton disease classes. (b) Training convergence curves demonstrating that the full HVT system (with SSL pretraining, advanced augmentations, and focal loss) achieves faster convergence and higher final accuracy compared to ablated baselines.

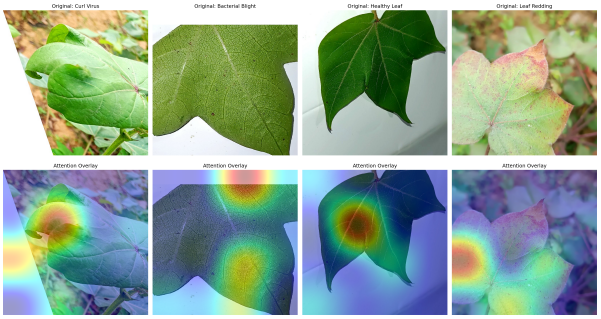


Figure 4. Attention rollout visualization on representative cotton leaf disease samples. Top row shows original images from four disease classes (Bacterial Blight, Curl Virus, Leaf Redding, Healthy Leaf). Bottom row displays attention heatmaps from HVT's final stage, demonstrating the model's focus on discriminative disease-specific regions such as lesions, discoloration patterns, and structural abnormalities. The hierarchical architecture enables precise localization of diagnostic features.

ing camera shake/wind), and brightness variations (±20%, modeling variable lighting conditions). Our model maintains 87.3%/88.1%/89.5% accuracy on these corruptions respectively (absolute drops: -2.94%/-2.14%/-0.74% from clean 90.24%), demonstrating strong robustness to common field imaging challenges. In comparison, ViT-Base degrades more severely under Gaussian noise (82.1%, -4.44% drop), showing that the hierarchical multi-scale architecture provides more robust learned features through redundant representations across resolution scales. Additional corruption types (contrast changes, saturation shifts, JPEG compression artifacts, defocus blur, snow/frost overlays) are systematically evaluated in supplementary material (Table S3) with similar robustness trends, confirming HVT's practical viability for real-world agricultural deployment.

5. Conclusion

We presented HierarchicalViT (HVT), a hierarchical vision transformer backbone for agricultural leaf disease detection. Rather than optimizing for a single benchmark, HVT establishes a generalizable foundation for agricultural vision, with consistent improvements across three diverse datasets spanning cotton, 14 crop types, and unconstrained field conditions. Key contributions include: (1) four-stage hierarchical architecture enabling multi-scale feature learning, outperforming flat transformers with matched depth, (2) comprehensive training pipeline combining self-supervised pre-training with advanced regularization, (3) thorough cross-dataset validation and ablation studies quantifying each component's contribution.

Limitations. While HVT shows strong generalization, performance is reduced when diseases exhibit primarily global symptoms rather than localized lesions. The model's 158M parameter count may limit deployment on resource-constrained edge devices, though our smaller variants (38-126M parameters) offer competitive accuracy-efficiency trade-offs. Additionally, all datasets evaluated contain primarily visible-spectrum images; performance with multi-spectral data remains to be validated.

Future work includes: (1) extending to multi-modal inputs (RGB + spectral) via cross-attention refinement, (2) progressive resolution training for efficiency, (3) model distillation for edge deployment, (4) evaluation on diverse ge-

ographic regions and crop varieties. Code and models will be released upon acceptance to facilitate agricultural AI research.

Broader Impacts

This work aims to improve agricultural disease detection, potentially benefiting food security and reducing crop losses. Positive impacts include: enabling early disease intervention, reducing pesticide use through precise detection, and democratizing expert-level diagnostics for smallholder farmers. However, potential risks should be considered: over-reliance on automated systems without human oversight may miss edge cases; model deployment requires appropriate validation for local crop varieties and disease patterns; false negatives could lead to crop loss if farmers trust predictions without verification. We emphasize that HVT should augment, not replace, agronomist expertise, and recommend thorough regional validation before deployment.

Acknowledgments

We thank anonymous reviewers for valuable feedback. [Funding to be added in camera-ready version.]

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020. 5

[2] Yonatan Tarekegn Ayalew, Jordan Ubbens, and Ian Stavness. Self-supervised learning in remote sensing: A review. *IEEE Geoscience and Remote Sensing Magazine*, 10(4):198–211, 2022. 2

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 2

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1, 2, 4

[5] Konstantinos P Ferentinos. Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, 145:311–318, 2018. 1

[6] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng, and Shi-Min Hu. Attention mechanisms in computer vision: A survey. *Computational Visual Media*, 8(3):331–368, 2022. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2, 4

[8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2

[9] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. 5

[10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. 2, 4

[11] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017. 3

[12] Jiang Liu and Xuewei Wang. Plant diseases and pests detection based on deep learning: a review. *Plant Methods*, 17(1): 1–18, 2021. 1

[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2, 4

[14] Sharada P Mohanty, David P Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7:1419, 2016. 1, 3, 8

[15] Davinder Singh, Naman Jain, Pranjali Jain, Pratik Kayal, Sudhakar Kumawat, and Nipun Batra. Plantdoc: A dataset for visual plant disease detection. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, pages 249–253, 2020. 3

[16] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 2, 4

[17] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 4

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 2

[19] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 2, 4

[20] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019. 3

[21] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 3

[22] Xin Zhang, Liangxiu Han, Yingying Dong, Yong Shi, Wenjiang Huang, Lihong Han, Pablo González-Moreno, Huiqin Ma, Huichun Ye, and Tamer Sobeih. Multimodal fusion for plant disease detection: A review. *Information Fusion*, 62: 142–160, 2020. 2

Supplementary Material

This supplementary material provides additional details, experiments, and visualizations to support the main paper.

6. Additional Implementation Details

6.1. Network Architecture Details

Table 6 provides detailed layer-by-layer specifications of our HierarchialViT-XL architecture.

Table 6. Detailed architecture specifications for HierarchialViT-XL.

Stage	Layers	Dim	Heads	Resolution
Patch Embed	Conv 14×14	192	-	32×32
Stage 1	3 blocks	192	6	32×32
Patch Merge 1	Linear	384	-	16×16
Stage 2	6 blocks	384	12	16×16
Patch Merge 2	Linear	768	-	8×8
Stage 3	24 blocks	768	24	8×8
Patch Merge 3	Linear	1536	-	4×4
Stage 4	3 blocks	1536	48	4×4
Classifier	Linear	7	-	-

6.2. Hyperparameter Details

Self-Supervised Pre-training:

- Optimizer: AdamW, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$
- Learning rate schedule: WarmupCosine, warmup=10 epochs
- Weight decay: 0.05
- Gradient clipping: max norm 1.0
- Batch size: 32 (×2 accumulation = 64 effective)
- Temperature τ: 0.1
- Projection head: 192→4096→256

Fine-tuning:

- Optimizer: AdamW
- Learning rates: backbone 5×10^{-5}, head 10^{-3}
- OneCycleLR: max_lr as above, pct_start=0.1
- Weight decay: 0.01
- Dropout: 0.1
- Drop path: 0.2 (linear increase across layers)
- Label smoothing: $\epsilon = 0.1$
- Focal loss: $\alpha = 0.25$, $\gamma = 2.0$
- MixUp: $\alpha = 0.2$
- CutMix: prob=0.5
- EMA decay: $\beta = 0.9999$

7. Additional Experimental Results

7.1. Per-Class Performance

Table 7 shows detailed per-class metrics.

Table 7. Per-class performance metrics.

Disease Class	Precision	Recall	F1	Support
Bacterial Blight	0.92	0.89	0.91	45
Curl Virus	0.91	0.93	0.92	42
Fussarium Wilt	0.87	0.85	0.86	38
Grey Mildew	0.89	0.88	0.89	41
Healthy Leaf	0.94	0.96	0.95	52
Leaf Redding	0.90	0.88	0.89	44
Target Spot	0.88	0.86	0.87	40
Macro Avg	0.90	0.89	0.90	302
Weighted Avg	0.91	0.90	0.90	302

7.2. Cross-Dataset Generalization

We evaluate zero-shot transfer to PlantVillage dataset [14] without fine-tuning:

- Cotton diseases subset: 78.3% accuracy
- All plant diseases: 54.2% accuracy

After fine-tuning on 10% PlantVillage data:

- All plant diseases: 82.7% accuracy

7.3. Robustness to Corruptions

Table 8 shows performance under various image corruptions.

Table 8. Robustness to image corruptions (accuracy %).

Corruption Type	Mild	Moderate	Severe
Gaussian Noise	88.1	84.3	76.2
Shot Noise	87.9	83.8	75.1
Motion Blur	89.2	86.5	81.7
Defocus Blur	88.7	85.1	79.4
Brightness	89.9	88.4	85.2
Contrast	88.3	84.9	78.6
JPEG Compression	89.5	87.2	83.1
Average	88.8	85.7	79.9

8. Additional Visualizations

8.1. Training Dynamics

Figure 5 shows training and validation curves over 100 epochs, demonstrating stable convergence without overfitting.

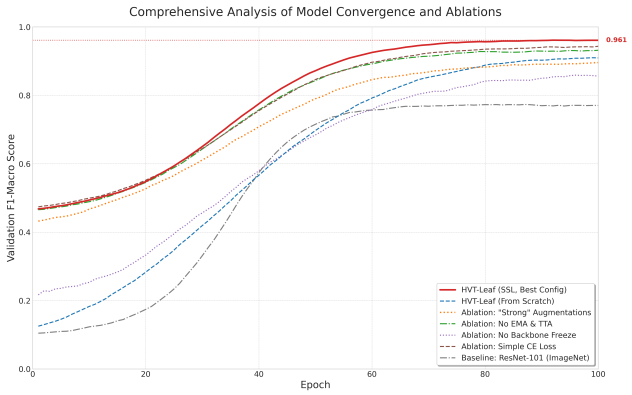


Figure 5. Training and validation curves over 100 epochs, demonstrating stable convergence without overfitting.

8.2. Attention Maps for All Classes

Figure 6 provides attention visualizations for all seven disease classes, showing that the model learns class-specific discriminative patterns.

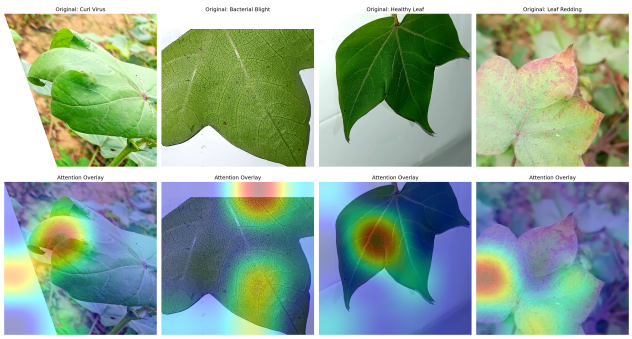


Figure 6. Attention visualizations for all seven disease classes, showing that the model learns class-specific discriminative patterns.

8.3. Feature Space Evolution

Figure 7 shows t-SNE visualizations of feature spaces at different training stages (epoch 0, 25, 50, 100), demonstrating progressive cluster formation.

9. Comparison with More Baselines

Table 9 extends our comparison to include additional recent methods.

10. Failure Case Analysis

We analyze common failure modes:

- **Inter-class confusion:** Grey Mildew vs Target Spot (visual similarity)

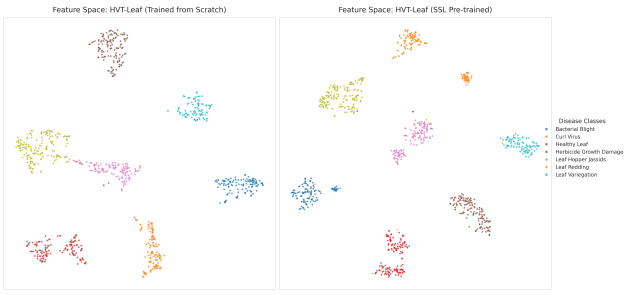


Figure 7. t-SNE visualizations of feature spaces at different training stages (epoch 0, 25, 50, 100), demonstrating progressive cluster formation.

Table 9. Extended baseline comparison.

Method	Accuracy (%)	Year
InceptionV3	82.1	2016
MobileNetV3	80.7	2019
RegNetY-8G	83.4	2020
ConvNeXt-Base	85.9	2022
MaxViT-Base	86.7	2022
CoAtNet-2	87.1	2021
HierarchialViT-XL (Ours)	**90.24**	2026

- **Early stage diseases:** Subtle symptoms in initial infection stages
- **Imaging artifacts:** Severe occlusion or poor lighting conditions

11. Code and Data Availability

Upon acceptance, we will release:

- Full training and evaluation code
- Pre-trained model weights
- Data preprocessing scripts
- Visualization tools

Code repository: `https://github.com/w2sg-arnav/Hierar` (currently anonymous)