HierarchicalViT for Agricultural Disease Detection Supplementary Material

Manoj Suryawanshi Sudha Gupta

1. Detailed Architecture Specifications

1.1. Complete Model Configuration

Table **??** provides the complete architecture specifications for HVT-XL.

Table 1. Complete HVT-XL Architecture Details

Stage	Resolution	Channels	Blocks	Heads	DropPath
1	32×32	192	3	6	0.0-0.075
2	16×16	384	6	12	0.075-0.15
3	8×8	768	24	24	0.15-0.3
4	4×4	1536	3	48	0.3

2. Mathematical Formulations

2.1. Stochastic Depth (DropPath)

The DropPath operation $\mathcal{D}(\cdot)$ applies stochastic depth regularization with drop probability p that increases linearly from 0 to 0.3 across layers:

$$\mathcal{D}(\mathbf{x}) = \begin{cases} \frac{\mathbf{x}}{1-p} \cdot \mathbf{b}, & \text{during training} \\ \mathbf{x}, & \text{during inference} \end{cases} \tag{1}$$

where $\mathbf{b} \sim \text{Bernoulli}(1-p)$ is a binary random variable. The scaling factor $1/(1-p)$ ensures that the expected value during training matches the deterministic inference pass.

The layer-wise drop probability is computed as:

$$p_l = p_{\text{max}} \cdot \frac{l}{L_{\text{total}}} \tag{2}$$

where l is the layer index, L_{total} is the total number of transformer blocks, and $p_{\text{max}} = 0.3$.

2.2. Feed-Forward Network

The feed-forward network $\mathcal{F}(\cdot)$ applies a two-layer MLP with GELU activation and expansion ratio $r = 4$:

$$\mathcal{F}(\mathbf{x}) = \mathbf{W}_2 \cdot \text{GELU}(\mathbf{W}_1 \mathbf{x}) \tag{3}$$

where:

- $\mathbf{W}_1 \in \mathbb{R}^{D \times rD}$ expands the dimension
- $\mathbf{W}_2 \in \mathbb{R}^{rD \times D}$ projects back to original dimension
- $\text{GELU}(x) = x \cdot \Phi(x)$ where $\Phi(x)$ is the Gaussian cumulative distribution function

The GELU activation can be approximated as:

$$\text{GELU}(x) \approx 0.5x \left(1 + \tanh\left[\sqrt{\frac{2}{\pi}}\left(x + 0.044715x^3\right)\right]\right) \tag{4}$$

2.3. Patch Merging Operation

The patch merging operation $\mathcal{M}(\cdot)$ reduces spatial resolution by concatenating 2×2 neighborhoods and projecting to higher dimension:

Given input $\mathbf{X}_s \in \mathbb{R}^{(H_s \times W_s) \times D_s}$ at stage s:

$$\mathbf{X}'_s = \text{Concat}(\mathbf{X}_s^{00}, \mathbf{X}_s^{01}, \mathbf{X}_s^{10}, \mathbf{X}_s^{11}) \in \mathbb{R}^{(H_s/2 \times W_s/2) \times 4D_s} \tag{5}$$

$$\mathbf{X}_{s+1} = \mathbf{W}_{\text{merge}} \mathbf{X}'_s \in \mathbb{R}^{(H_s/2 \times W_s/2) \times 2D_s} \tag{6}$$

where $\mathbf{W}_{\text{merge}} \in \mathbb{R}^{4D_s \times 2D_s}$ is a learned linear projection, and $\mathbf{X}_s^{ij}$ denotes the spatially shifted feature maps.

2.4. Self-Supervised Pre-training Loss

The complete NT-Xent (Normalized Temperature-scaled Cross Entropy) loss for SimCLR:

$$\mathcal{L}_{\text{SimCLR}} = -\frac{1}{2B} \sum_{i=1}^{2B} \left[\mathcal{L}_{\text{SimCLR}}^{(i,j(i))} + \mathcal{L}_{\text{SimCLR}}^{(j(i),i)}\right] \tag{7}$$

where $j(i)$ is the index of the positive pair for sample i, and:

$$\mathcal{L}_{\text{SimCLR}}^{(i,j)} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2B} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)} \tag{8}$$

The cosine similarity is computed as:

$$\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\|_2 \|\mathbf{v}\|_2} \tag{9}$$

Temperature $\tau = 0.5$ controls the concentration of the distribution.

2.5. Fine-tuning Training Objectives

Focal Loss. The focal loss with class weights:

$$\mathcal{L}_{\text{focal}} = -\sum_{c=1}^{C} \alpha_c (1 - p_c)^{\gamma} y_c \log(p_c) \tag{10}$$

where:
- $\alpha_c = 1/C$ for uniform class weighting (we use $\alpha_c = 1/7$)
- $\gamma = 2.0$ is the focusing parameter
- p_c is the predicted probability for class c
- $y_c \in \{0, 1\}$ is the ground truth

MixUp Augmentation. Sample mixing with Beta distribution:

$$\lambda \sim \text{Beta}(\alpha, \alpha), \quad \alpha = 0.2 \tag{11}$$

$$\tilde{\mathbf{x}} = \lambda \mathbf{x}_i + (1 - \lambda)\mathbf{x}_j \tag{12}$$

$$\tilde{\mathbf{y}} = \lambda \mathbf{y}_i + (1 - \lambda)\mathbf{y}_j \tag{13}$$

CutMix Augmentation. Regional replacement:

$$\tilde{\mathbf{x}} = \mathbf{M} \odot \mathbf{x}_i + (1 - \mathbf{M}) \odot \mathbf{x}_j \tag{14}$$

where $\mathbf{M} \in \{0, 1\}^{H \times W}$ is a binary mask indicating the region to replace. The mixing ratio λ is proportional to the area of the mask.

2.6. Learning Rate Schedules

WarmupCosine (Pre-training).

$$\eta_t = \begin{cases} \frac{t}{t_w}\eta_0, & t < t_w \\ \eta_0 \cdot \frac{1}{2}\left(1 + \cos\left(\pi \frac{t - t_w}{T - t_w}\right)\right), & t \geq t_w \end{cases} \tag{15}$$

where $t_w = 10$ epochs is the warmup period, $T = 80$ epochs is total training, and $\eta_0 = 5 \times 10^{-4}$.

OneCycleLR (Fine-tuning).

$$\eta_t = \begin{cases} \eta_{\min} + (\eta_{\max} - \eta_{\min})\frac{t}{t_{\text{warmup}}}, & t < t_{\text{warmup}} \\ \eta_{\max} - (\eta_{\max} - \eta_{\min})\frac{t - t_{\text{warmup}}}{T - t_{\text{warmup}}}, & t \geq t_{\text{warmup}} \end{cases} \tag{16}$$

with $\eta_{\max} = 0.1$, $\eta_{\min} = 10^{-5}$, and $t_{\text{warmup}} = 0.1T$.

Exponential Moving Average (EMA).

$$\theta_{\text{EMA}}^{(t)} = \beta \theta_{\text{EMA}}^{(t-1)} + (1 - \beta)\theta^{(t)} \tag{17}$$

where $\beta = 0.9999$ provides smoothing over approximately 10,000 gradient updates.

3. Additional Experimental Results

3.1. Per-Class Performance Analysis

Table **??** shows detailed per-class precision, recall, and F1 scores.

Table 2. Per-Class Performance Metrics

Disease Class	Precision	Recall	F1
Healthy	95.2	96.8	96.0
Bacterial Blight	88.7	87.3	88.0
Curl Virus	91.4	92.1	91.7
Fussarium Wilt	89.8	88.5	89.1
Grey Mildew	87.2	86.9	87.0
Leaf Redding	93.5	94.2	93.8
Target Spot	86.3	85.7	86.0
Average	**90.3**	**90.2**	**90.2**

Table 3. Complete ImageNet-C Corruption Results (Accuracy %)

Corruption	ResNet-101	ViT-B	Swin-B	HVT
Gaussian Noise	68.3	71.2	74.5	**78.9**
Shot Noise	69.1	72.0	75.2	**79.3**
Impulse Noise	67.8	70.5	73.8	**77.6**
Defocus Blur	74.2	76.8	79.1	**82.4**
Glass Blur	70.5	73.4	76.2	**79.8**
Motion Blur	71.8	74.5	77.3	**80.7**
Zoom Blur	72.3	75.1	78.0	**81.2**
Snow	73.5	76.2	78.9	**81.8**
Frost	72.1	74.8	77.5	**80.4**
Fog	75.8	78.3	80.6	**83.1**
Brightness	79.2	81.5	83.2	**85.6**
Contrast	76.4	78.9	81.3	**84.0**
Elastic Transform	73.9	76.5	79.2	**82.1**
Pixelate	74.6	77.3	79.8	**82.5**
JPEG Compression	77.1	79.6	82.0	**84.8**
Average	**73.1**	**75.7**	**78.4**	**81.6**

3.2. Complete ImageNet-C Robustness Results

We evaluate robustness on all 15 ImageNet-C corruption types. Table **??** shows complete results.

3.3. Training Time Analysis

Table **??** provides detailed training time breakdown.

Table 4. Training Time Breakdown

Phase	Epochs	Time (NVIDIA T4)
SSL Pre-training	80	12h (avg. 9 min/epoch)
Fine-tuning	100	8h (avg. 4.8 min/epoch)
Hyperparameter Search	-	24h (12 configs)
Baseline Training	-	32h (4 models × 8h)
Total	-	**76h**

4. Implementation Details

4.1. Data Augmentation Pipeline

Pre-training (SimCLR):

- Random resized crop: scale (0.2, 1.0), ratio (0.75, 1.33)
- Color jitter: brightness=0.4, contrast=0.4, saturation=0.4, hue=0.1
- Random grayscale: p=0.2
- Gaussian blur: kernel size=23, $\sigma \in [0.1, 2.0]$
- Random horizontal flip: p=0.5

Fine-tuning:

- Random resized crop: scale (0.8, 1.0)
- Random horizontal flip: p=0.5
- Random vertical flip: p=0.5
- Random rotation: degrees=(-15, 15)
- Color jitter: brightness=0.2, contrast=0.2
- MixUp: p=0.2, $\alpha = 0.2$
- CutMix: p=0.5, $\alpha = 1.0$

4.2. Optimizer Configuration

Pre-training (AdamW):

- Initial learning rate: 5×10^{-4}
- Weight decay: 0.05
- Betas: (0.9, 0.999)
- Epsilon: 10^{-8}
- Gradient clipping: max norm = 1.0

Fine-tuning (AdamW):

- Initial learning rate: 0.1 (with OneCycleLR)
- Weight decay: 10^{-4}
- Betas: (0.9, 0.999)
- Layer-wise learning rate decay: 0.65
- Gradient clipping: max norm = 5.0

5. Qualitative Results

5.1. Additional Attention Visualizations

Figure **??** shows attention maps for additional disease classes demonstrating the hierarchical multi-scale processing across all four stages.

5.2. Failure Case Analysis

Figure **??** presents challenging cases where all models struggle, including:

- Extreme occlusion (¿70% leaf area hidden)
- Very early-stage disease (minimal symptoms)
- Multiple co-occurring diseases
- Poor lighting conditions
- Severe image blur or motion artifacts

While HVT shows better robustness than baselines, these cases highlight limitations and areas for future improvement.

6. Dataset Details

6.1. Cotton Leaf Disease Dataset Statistics

Table 5. Dataset Statistics

Class	Train	Val	Test	Total
Healthy	700	100	100	900
Bacterial Blight	490	70	70	630
Curl Virus	420	60	60	540
Fussarium Wilt	350	50	50	450
Grey Mildew	280	40	40	360
Leaf Redding	210	30	30	270
Target Spot	245	35	35	315
Unlabeled (SSL)	3000	-	-	3000
Total	**2695**	**385**	**385**	**7465**

6.2. Data Collection Protocol

Images were collected from cotton fields in three different geographic regions over the 2023-2024 growing season. Expert plant pathologists provided disease annotations. Images were captured using smartphone cameras (12MP resolution) under natural lighting conditions at various times of day. The dataset will be released upon paper acceptance with detailed annotation guidelines and metadata.