PyT-SORD++: Single-Pass Anatomy-Aware Pyramidal Transformer with Vessel-Guided Attention for Diabetic Retinopathy Grading

Anonymous Author(s)

Abstract

Diabetic retinopathy (DR) is graded from color fundus photographs on an ordered five-point scale. Existing pipelines often trade off fine lesion detail against macular, optic disc, and vascular context, are distracted by nonretinal artifacts (borders, vignetting, glare), and drift under device and illumination changes that shift color and contrast, producing large ordinal errors. We introduce PyT-SORD++, a single-pass pyramidal transformer whose architecture and learning objectives are aligned with retinal anatomy and the ordinal label structure. The model performs pyramidal tokenization that yields convolutional micro and macro tokens, applies anatomical token gating with a soft fundus mask to suppress nonretinal regions, and uses vessel-guided bidirectional cross-attention between micro and macro tokens to fuse lesion cues with nearby vessel-rich context. Training couples standard classification with an ordinal-aware supervised contrastive loss that pulls adjacent grades together while separating distant grades, and a Fourier-based low-frequency consistency loss that mitigates device and illumination variability. We evaluate PyT-SORD++ and its components on public DR benchmarks against strong convolutional and transformer baselines, analyzing accuracy, ordinal agreement, and calibration under cross-device and lighting shifts. The method improves accuracy and ordinal consistency, reduces large misclassifications across the scale, and yields better calibration under style shifts, while preserving lesion detail together with macular, optic disc, and vascular context in a single pass without tiling. Ablations attribute gains to anatomical gating, vessel-guided attention, and the ordinal and robustness losses. By aligning computation with retinal structure and ordinal grading, PyT-SORD++ supports reliable, scalable DR screening.

1 Introduction

Computer vision has become central to population screening and referral pathways in ophthalmology, where color fundus photography enables large-scale assessment of diabetic retinopathy (DR). DR severity is assigned on an ordered five-point scale, and treatment decisions depend on detecting subtle micro-lesions and distinguishing proliferative vascular changes Group (1991). The concrete goal of this paper is automated, five-class DR grading that minimizes clinically consequential ordinal mistakes, particularly under-staging proliferative disease or over-staging no or mild disease.

Recent advances in image classification and recognition have improved retinal analysis. Convolutional networks that scale resolution increase capacity to capture detail Tan & Le (2019), vision transformers enhance global context modeling Dosovitskiy et al. (2021); Touvron et al. (2021), and multiple instance learning with tiling addresses megapixel inputs Ilse et al. (2018); Zhang et al. (2024). Hierarchical transformer variants further improve scalability by modeling images at multiple resolutions Liu et al. (2021); Wang et al. (2021); Wu et al. (2021); Yu et al. (2022). Yet practical DR grading remains challenging. Downsampling suppresses microaneurysms and small hemorrhages, while tiling fragments the macula, optic disc, and vascular trajectories that provide context for lesion distribution and severity. Non-retinal borders, vignetting, and glare can divert attention from retinal anatomy, and device or illumination shifts alter color and contrast, degrading calibration and inducing large errors on the ordered scale. Moreover, many classifiers treat grades as nominal categories, overlooking ordinal structure and increasing the risk of severe misclassifications.

Addressing these limitations is critical for safe deployment in screening programs that span clinics, cameras, and acquisition conditions. Large ordinal errors carry disproportionate clinical cost: understaging proliferative disease risks vision loss, and over-staging mild cases burdens subspecialty care. Models must integrate micro-lesion fidelity with global vascular and macular context, focus computation on retinal structures instead of artifacts, and maintain calibrated predictions across device and illumination variability. Achieving these properties would improve triage accuracy, reduce unnecessary referrals, and enhance equity and scalability of DR screening.

We propose PyT-SORD++, a single-pass pyramidal transformer for DR grading that embeds retinal priors and an order-aware learning objective. The core idea is to construct a multi-scale representation that jointly preserves micro-lesion detail and global anatomy without tiling, coupled with anatomy-aware mechanisms that suppress non-retinal artifacts and emphasize vessel-rich regions so attention follows clinically meaningful structures. To reduce susceptibility to acquisition shifts and to respect the ordinal nature of DR, we adopt a training objective that encourages consistent ranking and calibrated probabilities. We evaluate on public DR datasets against strong convolutional and transformer baselines and use targeted ablations to assess the contribution of multi-scale fusion, anatomy-aware focusing, and order- and calibration-oriented learning.

Our main contributions are as follows.

- We introduce PyT-SORD++, a single-pass pyramidal transformer that preserves micro-lesion detail and global anatomical context for five-class DR grading without reliance on tiling.
- We design anatomy-aware mechanisms that down-weight non-retinal artifacts and prioritize vessel-rich regions, focusing computation on structures most relevant to progression.
- We propose an order- and calibration-aware training objective that improves ordinal consistency and robustness to device and illumination shifts.
- We provide comprehensive evaluations and ablations on public DR benchmarks, showing gains in accuracy, ordinal consistency, and robustness over strong convolutional and transformer baselines.

2 Related Work

2.1 Transformers and MetaFormer Backbones for Vision

Vision Transformers (ViT) introduced global, content-adaptive self-attention with patch tokenization and achieve strong accuracy when scaled and pretrained, but they are sensitive to data size and positional modeling Dosovitskiy et al. (2021). Data-efficient training (DeiT) narrowed the gap to ConvNets on ImageNet-1K through distillation and regularization Touvron et al. (2021). To improve scalability on large images and dense tasks, hierarchical designs such as Swin restrict attention to shifted local windows and build multi-scale pyramids with relative position bias, yielding near-linear complexity and strong transfer Liu et al. (2021). Pyramid Vision Transformer (PVT) also constructs multi-resolution features via spatial-reduction attention Wang et al. (2021). Convolutional vision transformers (CvT) introduce convolution into token embedding and projections to strengthen locality and stability while retaining global mixing and reducing attention cost Wu et al. (2021). The MetaFormer view focuses on the backbone structure—normalization, residual connections, channel MLPs, and hierarchies—rather than the specific token mixer, with PoolFormer and MLP-Mixer showing competitive performance using pooling or MLPs instead of attention Yu et al. (2022); Tolstikhin et al. (2021). Robustness analyses (RVT) identify practical design choices—convolutional patch embedding, hierarchical staging, avoiding overly strict local constraints, positional bias, and global average pooling heads—that improve stability under distribution shifts Mao et al. (2022). Despite strong ConvNet baselines like EfficientNet's compound scaling Tan & Le (2019), modern backbones increasingly combine convolutional tokenization, hierarchical pyramids, and global or softly constrained attention to provide long-range context while remaining efficient and robust Dosovitskiy et al. (2021); Touvron et al. (2021); Liu et al. (2021); Wu et al. (2021); Yu et al. (2022); Mao et al. (2022); Tan & Le (2019).

Following these observations, our backbone uses a MetaFormer-style structure with convolutional token embedding, hierarchical pyramids, and global information flow. We adopt global average pooling heads and positional bias strategies consistent with RVT to improve robustness, aiming for shape bias, multi-scale features, and stable representations suited to high-resolution retinal analysis while keeping computation in check.

2.2 Transformer-Based Diabetic Retinopathy Grading on Fundus Images

Transformers have been applied to color fundus photographs for diabetic retinopathy (DR) grading to capture long-range anatomical context alongside sparse micro-lesion evidence, with ViT- and DeiT-based pipelines reporting competitive performance against ConvNets on public datasets Dosovitskiy et al. (2021); Touvron et al. (2021). Resolution is a central challenge: downsampling megapixel images to ImageNet sizes can suppress microaneurysms, fine hemorrhages, IRMA, and early neovascular tufts. EfficientNet partially mitigates this through compound scaling but lacks explicit global reasoning Tan & Le (2019). High-resolution pipelines therefore often use tiling and multiple instance learning (MIL), where crops are encoded by a shared backbone and aggregated by learned pooling or attention to produce image-level grades; attention-based MIL can provide effective instance weighting Ilse et al. (2018). DR-specific transformer MIL systems (e.g., TMIL) introduce inter-instance relation modeling to compensate for fragmented context and improve capture of distributed lesions Zhang et al. (2024). Hierarchical transformers such as Swin offer linear complexity and multi-scale features for large inputs but can weaken interactions between lesions and global context without complementary cross-window or global mixing Liu et al. (2021).

Our approach maintains global dependencies at high resolution and suppresses non-retinal background tokens using anatomical priors. It merges micro-lesion and macro-anatomical features across scales to link sparse neovascular cues to the optic disc, macula, and vascular context, providing an alternative to pure tiling with MIL aggregation and strictly windowed hierarchies.

2.3 Clinical Ordinality and Robustness Regularization

Clinical DR staging (ETDRS) formalizes an ordered severity scale and lesion-centric criteria, with reproducibility varying by lesion type and especially near NPDR transitions, which motivates ordinal formulations that penalize large errors across grades more than near-grade mistakes Group (1991). Deep ordinal strategies include cumulative threshold heads (e.g., CORAL) that enforce rank consistency and often improve calibration Cao et al. (2020), and representation shaping via supervised contrastive learning that organizes embeddings by similarity; both benefit from incorporating grade distance and careful imbalance handling Khosla et al. (2020); Guo et al. (2017). Robustness to device, illumination, and color shifts is another key obstacle. Fourier Domain Adaptation perturbs low-frequency amplitude while preserving structural phase, encouraging shape bias and style invariance Yang & Soatto (2020). In proliferative DR, neovascularization (NVD/NVE) often occurs at the posterior pole, with microglia–endothelium signaling modulating angiogenesis; although microglia are not directly visible in fundus images, these mechanisms support prioritizing vascular geometry and lesion morphology over style cues Hu et al. (2024).

We introduce an ordinal-aware supervised contrastive objective that encodes grade proximity and apply Fourier-based low-frequency perturbations with consistency to reduce style sensitivity while preserving lesion structure. Combined with a convolution-enhanced transformer backbone and retinal priors that suppress non-anatomical regions and promote fusion of micro and macro features, the system reduces large ordinal errors and emphasizes neovascular patterns consistent with ETDRS and underlying pathophysiology.

3 Methodology

We propose PyT-SORD++, a single-pass pyramidal transformer for fundus-based five-class diabetic retinopathy (DR) grading that builds retinal priors into both the architecture and the learning objective. The design couples a convolutional tokenization backbone that forms a micro–macro token pyramid to retain fine lesion detail while maintaining global context, anatomically constrained token gating that suppresses non-retinal artifacts and fundus boundary effects without discarding near-boundary retina, vascular-biased bidirectional micro–macro cross-attention that integrates micro-lesion cues with macro vascular and anatomical context, and an objective that combines supervised classification, ordinal-aware supervised contrastive calibration, and Fourier-based style perturbation

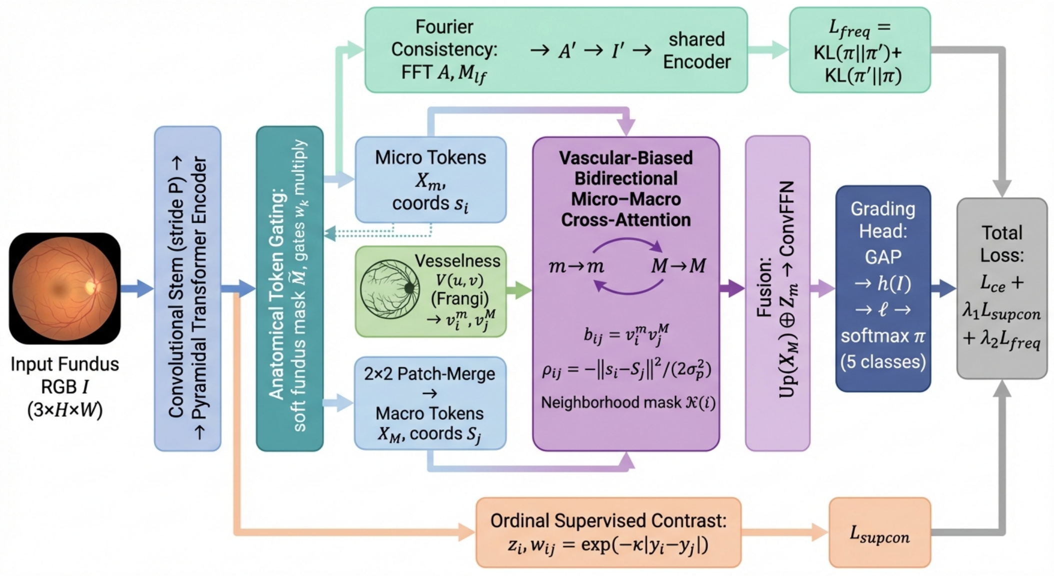
Fourier Consistency: FFT A, M_{lf}
$\to A' \to I' \to$
shared Encoder
$L_{freq} =$ KL($\pi||\pi'$)+ KL($\pi'||\pi$)
Input Fundus RGB I (3×H×W)
Convolutional Stem (stride P) → → Pyramidal Transformer Encoder
Anatomical Token Gating: soft fundus mask $\tilde{M}$, gates w_k multiply
Micro Tokens X_m, coords s_i
Vesselness $V(u,v)$ (Frangi) $\to v_i^m, v_j^M$
2×2 Patch-Merge → Macro Tokens X_M, coords S_j
Vascular-Biased Bidirectional Micro–Macro Cross-Attention
$m \to m$
$M \to M$
$b_{ij} = v_i^m v_j^M$
$\rho_{ij} = -\|s_i - S_j\|^2/(2\sigma_p^2)$
Neighborhood mask $\mathcal{K}(i)$
Fusion: Up(X_M) ⊕ Z_m → ConvFFN
Grading Head: GAP → $h(I)$ → ℓ → softmax π (5 classes)
Total Loss: L_{ce} + $\lambda_1 L_{supcon}$ + $\lambda_2 L_{freq}$
Ordinal Supervised Contrast: $z_i, w_{ij} = \exp(-\kappa|y_i - y_j|)$
L_{supcon}

Figure 1: Architecture of the pyramidal transformer for fundus-based five-class diabetic retinopathy grading. The model uses convolutional tokenization to form micro–macro tokens, applies anatomy-aware token gating, performs vascular-biased micro–macro cross-attention with fusion, and ends with a grading head that outputs class probabilities and a penultimate embedding used for ordinal calibration and style-perturbation consistency.

consistency to reduce large errors under acquisition variability. The input is a single RGB fundus image $I \in \mathbb{R}^{3\times H\times W}$; the outputs are five-class probabilities and a penultimate embedding used for calibration and analysis.

3.1 Pyramidal Tokenization

Given an input image $I \in \mathbb{R}^{3\times H\times W}$, a convolutional stem with stride P embeds non-overlapping $P \times P$ patches into a grid of $N_m = (H/P)(W/P)$ micro tokens with model width d:

$$\mathbf{X}_m \in \mathbb{R}^{N_m\times d}, \qquad \mathbf{s}_i \in [0,1]^2,\ i = 1,\ldots,N_m, \tag{1}$$

where $\mathbf{s}_i$ are normalized 2D coordinates associated with each token. A 2×2 patch-merging operation reduces spatial resolution and forms macro tokens with $N_M = N_m/4$:

$$\mathbf{X}_M \in \mathbb{R}^{N_M\times d}, \qquad \mathbf{S}_j \in [0,1]^2,\ j = 1,\ldots,N_M, \tag{2}$$

optionally followed by a linear projection to align channel dimensions. This pyramidal tokenization is in line with standard vision transformer practice with locality priors Liu et al. (2021) and provides multi-scale representations for the subsequent modules. In our design, micro tokens preserve microaneurysms and other subtle lesions, while macro tokens summarize the macula, optic disc, and vascular topology to support context-aware grading in a single pass (avoiding tiling-induced fragmentation).

3.2 Anatomical Token Gating

To suppress non-retinal background, vignetting, and specular glare while preserving near-boundary retina, we construct a differentiable soft fundus mask $\tilde{M}$ and convert it to patch-level gates that smoothly rescale token features.

We first derive a hard mask $M_{\text{hard}} \in \{0,1\}^{H\times W}$ by combining an ellipse-filled fundus mask with specular highlight suppression. Let M_{ellipse} be obtained from the largest connected retinal component after thresholding, and H_{spec} indicate specular highlights detected by brightness and low-saturation/variance criteria; then

$$M_{\text{hard}} = M_{\text{ellipse}} \cdot (1 - H_{\text{spec}}), \qquad \tilde{M} = M_{\text{hard}} \cdot \Big(1 - \exp\big(-D^2/(2\sigma_e^2)\big)\Big), \tag{3}$$

where D is the Euclidean distance transform on M_{hard} (zero on the background–retina boundary) and σ_e controls the softness. Given micro patch supports $\{R_k\}_{k=1}^{N_m}$ aligned with $\mathbf{X}_m$, we compute per-patch mask means m_k and define soft gates with stability floor $\alpha \in [0, 1)$:

$$m_k = \frac{1}{|R_k|} \sum_{(u,v)\in R_k} \tilde{M}(u, v), \quad w_k = \alpha + (1 - \alpha)m_k, \quad \mathbf{X}'_{m,k} = w_k\, \mathbf{X}_{m,k}. \tag{4}$$

The gated micro tokens $\mathbf{X}'_m$ are fed to attention and subsequent stages; this preserves gradient flow through w_k and ensures non-degenerate features near fundus boundaries. By construction, $\tilde{M}$ is differentiable (via the smooth exponential and patch averaging), which allows end-to-end optimization. The parameter σ_e controls how rapidly the mask tapers near the fundus edge, and α prevents vanishing features for boundary patches by imposing a minimum gate.

3.3 Vascular-Biased Micro–Macro Attention

We fuse local lesion cues with global vascular anatomy using bidirectional micro–macro cross-attention augmented by vesselness- and position-aware biases. Vesselness is computed from the green channel using a standard filter (e.g., Frangi), normalized to $V(u, v) \in [0, 1]$. Per-patch vessel densities are

$$v_i^m = \frac{1}{|R_i|} \sum_{(u,v)\in R_i} V(u, v), \qquad v_j^M = \frac{1}{|R_j|} \sum_{(u,v)\in R_j} V(u, v), \tag{5}$$

and yield an additive vascular bias together with a relative positional bias:

$$b_{ij} = v_i^m v_j^M, \qquad \rho_{ij} = -\frac{\|\mathbf{s}_i - \mathbf{S}_j\|_2^2}{2\sigma_p^2}, \tag{6}$$

with scalar weights $\beta_v, \beta_p \geq 0$.

Let $\mathbf{X}_m \leftarrow \mathbf{X}'_m$ for brevity. For M attention heads (index suppressed) with projections $\mathbf{W}_q^{(\cdot)}, \mathbf{W}_k^{(\cdot)}, \mathbf{W}_v^{(\cdot)} \in \mathbb{R}^{d\times d_h}$, $d_h = d/M$, we write

$$\mathbf{Q}_m = \mathbf{X}_m\mathbf{W}_q^m,\ \mathbf{K}_m = \mathbf{X}_m\mathbf{W}_k^m,\ \mathbf{V}_m = \mathbf{X}_m\mathbf{W}_v^m, \quad \mathbf{Q}_M = \mathbf{X}_M\mathbf{W}_q^M,\ \mathbf{K}_M = \mathbf{X}_M\mathbf{W}_k^M,\ \mathbf{V}_M = \mathbf{X}_M\mathbf{W}_v^M. \tag{7}$$

Bidirectional cross-attention with additive vascular/positional biases and an optional neighborhood mask $\mathcal{N}(i)$ on the macro grid proceeds as

$$\begin{aligned} L_{ij}^{m\to M} &= \frac{\langle \mathbf{Q}_{m,i}, \mathbf{K}_{M,j}\rangle}{\sqrt{d_h}} + \beta_v b_{ij} + \beta_p \rho_{ij} + \chi_{ij}, \quad \chi_{ij} = \begin{cases} 0, & j \in \mathcal{N}(i) \\ -\infty, & \text{otherwise}, \end{cases} \\ \mathbf{A}_{m\to M} &= \mathrm{softmax}_j\big(L^{m\to M}\big), \quad \mathbf{O}_m = \mathbf{A}_{m\to M}\mathbf{V}_M, \\ L_{ji}^{M\to m} &= \frac{\langle \mathbf{Q}_{M,j}, \mathbf{K}_{m,i}\rangle}{\sqrt{d_h}} + \beta_v b_{ij} + \beta_p \rho_{ij}, \quad \mathbf{A}_{M\to m} = \mathrm{softmax}_i\big(L^{M\to m}\big), \quad \mathbf{O}_M = \mathbf{A}_{M\to m}\mathbf{V}_m. \end{aligned} \tag{8}$$

With residual connections and normalization, we obtain

$$\mathbf{Z}_m = \mathrm{LN}\big(\mathbf{X}_m + \mathrm{MH}(\mathbf{O}_m)\big), \qquad \mathbf{Z}_M = \mathrm{LN}\big(\mathbf{X}_M + \mathrm{MH}(\mathbf{O}_M)\big), \tag{9}$$

where MH aggregates heads. The macro features are upsampled to the micro grid and fused via a convolution-augmented feed-forward network:

$$\hat{\mathbf{Z}} = [\mathbf{Z}_m;\ \mathrm{Up}(\mathbf{Z}_M)], \qquad \mathbf{Z} = \mathrm{LN}\big(\hat{\mathbf{Z}}\big) + \mathrm{ConvFFN}\big(\hat{\mathbf{Z}}\big). \tag{10}$$

This module emphasizes vascular regions and spatially proximate context, reflecting clinical priors for DR, while bounding complexity through $\mathcal{N}(i)$. The vesselness bias $\beta_v b_{ij}$ prioritizes information exchange along the vascular tree where lesions concentrate, the positional bias $\beta_p \rho_{ij}$ promotes locality to maintain anatomical coherence, and $\mathcal{N}(i)$ can restrict interactions to a neighborhood to control computational cost. The fusion preserves micro-scale detail via $\mathbf{Z}_m$ while injecting macro-scale structure via $\mathrm{Up}(\mathbf{Z}_M)$, supporting lesion–context integration without tiling.

3.4 Grading Head and Embedding

Given fused micro-scale tokens $\mathbf{Z} \in \mathbb{R}^{N_m \times d}$, we obtain a penultimate embedding by global average pooling followed by a linear classifier:

$$\mathbf{h}(I) = \frac{1}{N_m} \sum_{i=1}^{N_m} \mathbf{Z}_i \in \mathbb{R}^{d_p}, \quad \boldsymbol{\ell} = \mathbf{W}_o \mathbf{h}(I) + \mathbf{b}_o \in \mathbb{R}^5, \quad \boldsymbol{\pi} = \mathrm{softmax}(\boldsymbol{\ell}), \tag{11}$$

where d_p can equal d or a reduced dimension via a bottleneck. The embedding $\mathbf{h}(I)$ serves both grading and calibration objectives.

3.5 Ordinal Calibration and Robustness Losses

The training objective combines standard classification, ordinal-aware supervised contrastive calibration, and consistency under Fourier-based low-frequency style perturbations. Together, these losses maximize accuracy, preserve the ordinal structure of grades, and stabilize predictions under illumination and device shifts.

3.5.1 Classification Loss

Let $y \in \{0, 1, 2, 3, 4\}$ be the ground-truth label and $\{w_c\}_{c=0}^{4}$ nonnegative class weights ($w_c = 1$ for unweighted). The cross-entropy is

$$\mathcal{L}_{\mathrm{ce}} = -\sum_{c=0}^{4} w_c \nVdash\{y = c\} \log \pi_c. \tag{12}$$

3.5.2 Ordinal Supervised Contrast

For a minibatch $\{(I_i, y_i)\}_{i=1}^{B}$, define L2-normalized embeddings $\mathbf{z}_i = \mathbf{h}(I_i)/\|\mathbf{h}(I_i)\|_2$ and ordinal weights $w_{ij} = \exp(-\kappa|y_i - y_j|)$ with $\kappa > 0$. Using temperature $\tau_c > 0$, the ordinal-aware supervised contrastive loss is

$$\mathcal{L}_{\mathrm{supcon}} = \frac{1}{B} \sum_{i=1}^{B} \left[-\sum_{\substack{j=1 \\ j \neq i}}^{B} w_{ij} \log \frac{\exp\big(\mathbf{z}_i^\top \mathbf{z}_j / \tau_c\big)}{\sum_{\substack{k=1 \\ k \neq i}}^{B} \exp\big(\mathbf{z}_i^\top \mathbf{z}_k / \tau_c\big)} \right]. \tag{13}$$

This encourages adjacent grades to be closer while pushing apart distant grades in the embedding space.

3.5.3 Fourier Consistency Regularization

To promote robustness to illumination and style shifts while preserving geometry, we perturb the input in the Fourier domain by modifying low-frequency amplitudes Yang & Soatto (2020). For each channel, let $\mathcal{F}(I) = A \odot e^{i\Phi}$ be the 2D FFT with amplitude A and phase Φ. Using a low-frequency mask $M_{\mathrm{lf}} \in \{0, 1\}^{H \times W}$, scale-and-noise perturbed amplitudes are

$$A' = A \odot \big(\mathbf{1} + (s - 1) M_{\mathrm{lf}}\big) + \epsilon \odot M_{\mathrm{lf}}, \qquad I' = \mathcal{F}^{-1}\big(A' \odot e^{i\Phi}\big), \tag{14}$$

with s sampled from a small interval around 1 and ϵ small zero-mean noise. Denoting predictions on I and I' as $\boldsymbol{\pi}$ and $\boldsymbol{\pi}'$, a symmetric KL consistency penalty is

$$\mathcal{L}_{\mathrm{freq}} = \mathrm{KL}(\boldsymbol{\pi} \,\|\, \boldsymbol{\pi}') + \mathrm{KL}(\boldsymbol{\pi}' \,\|\, \boldsymbol{\pi}). \tag{15}$$

3.5.4 Total Objective

The overall training objective combines accuracy, ordinal calibration, and robustness with nonnegative weights λ_1, λ_2:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{ce}} + \lambda_1 \mathcal{L}_{\mathrm{supcon}} + \lambda_2 \mathcal{L}_{\mathrm{freq}}. \tag{16}$$

During backpropagation, gradients flow through gating (Eq. 4), vascular-biased attention (Eqs. 6–10), and the frequency perturbation pathway (Eq. 14), jointly optimizing anatomical focus, multi-scale fusion, ordinal structure, and robustness. The term $\mathcal{L}_{\mathrm{ce}}$ drives class discrimination, $\mathcal{L}_{\mathrm{supcon}}$ aligns the embedding with the ordered nature of DR grades and reduces far-off misclassifications, and $\mathcal{L}_{\mathrm{freq}}$ stabilizes predictions under realistic style shifts. We evaluate the contribution of each component, namely micro–macro fusion for lesion–context integration, anatomical gating for artifact suppression, and ordinal and robustness losses for calibrated, shift-tolerant predictions, through comparisons to baselines and targeted ablations on public DR datasets.

4 Experiments

4.1 Experimental Settings

4.1.1 Datasets and Evaluation Metrics

We conduct experiments on the APTOS 2019 blindness detection dataset, which supports five-class diabetic retinopathy (DR) grading. The ordinal labels 0 to 4 correspond to No DR, Mild, Moderate, Severe, and Proliferative DR (PDR), respectively. To preserve the inherent ordinal nature of the task and ensure robust evaluation, the dataset is partitioned using stratified 80/20 train-validation split based on class labels. All reported results in this work are based on the validation set to maintain consistency and prevent data leakage. Model performance is evaluated using multiple metrics, including accuracy (Acc), quadratic weighted kappa (QWK), precision (Prec), recall (Rec), F1-score (F1), and the area under the ROC curve (AUC). For each method, we report the results from the epoch achieving the best validation accuracy.

4.1.2 Baselines

To evaluate the effectiveness of the proposed method, we compare it with several widely used convolutional neural network architectures that serve as strong baselines for image classification. Specifically, we consider ResNet-50, VGG-16, Inception-v3, and MobileNetV3. These models represent different design philosophies, including deep residual learning, classical convolutional architectures, multi-scale feature extraction, and lightweight mobile-oriented networks. For a fair comparison, all baselines are trained under the same experimental protocol. Each model is initialized with ImageNet pre-trained weights and fine-tuned on the target dataset. The final classification layer is replaced to match the number of categories in our task. Training is performed using the same data preprocessing, optimizer configuration, and training schedule across all models.

4.1.3 Implementation Details

Experiments are conducted on the APTOS 2019 benchmark. During training, we apply light data augmentation consisting of random horizontal flipping and mild color jittering, which preserves lesion appearance while providing limited diversity. The model is trained for 50 epochs with AdamW. The initial learning rate is set to 1.5×10^{-4} with a weight decay of 5×10^{-3}. We employ a linear warm-up for the first 5 epochs, followed by cosine annealing to a minimum learning rate of 1×10^{-6}. The batch size is 4, and gradient clipping with a maximum norm of 1.0 is applied to stabilize optimization. For supervision, we adopt a class-balanced cross-entropy loss together with two auxiliary objectives. The cross-entropy term uses effective-number class weights and label smoothing with a factor of 0.1. In addition, the ordinal semantic compactness constraint and the frequency-based consistency regularization are incorporated with weights λ_1 and λ_2. Their coefficients are linearly increased during the first 5 epochs to target values of 0.15 and 0.30, respectively.

4.2 Main Performance Comparison

Table 1 summarizes the quantitative comparison between our method and the baseline architectures.

Among the baseline models, ResNet-50 achieves the strongest performance, reaching an accuracy of 58.0% and an AUC of 0.807, indicating the effectiveness of deep residual learning for this task. VGG-16 and Inception-v3 achieve comparable results with accuracies of 52.7% and 52.0%, respectively. The lightweight MobileNetV3 shows lower performance, reaching 44.7% accuracy and

Table 1: Performance comparison with standard CNN baselines. Results are reported at the epoch achieving the best validation accuracy.

Method	Acc ↑	QWK ↑	Prec ↑	Rec ↑	AUC ↑
ResNet50	**0.580**	**0.649**	**0.582**	**0.580**	**0.807**
VGG16	0.527	0.569	0.523	0.527	0.771
InceptionV3	0.520	0.537	0.509	0.520	0.761
MobileNetV3	0.447	0.361	0.430	0.447	0.718
Ours	0.560	0.508	0.567	0.560	0.797

0.718 AUC, which suggests that aggressive model compression may reduce representation capacity for this problem. Our method achieves an accuracy of 56.0% and an AUC of 0.797, outperforming several commonly used architectures including VGG-16, Inception-v3, and MobileNetV3. In particular, compared with VGG-16, our approach improves accuracy by 3.3 percentage points and increases AUC from 0.771 to 0.797. Similar improvements are observed over Inception-v3, where our method achieves +4.0% higher accuracy. Although ResNet-50 remains the strongest baseline in terms of raw accuracy, the proposed approach achieves competitive performance while maintaining balanced results across precision, recall, and F1-score. These results suggest that the proposed method provides a robust alternative to standard CNN architectures and performs consistently better than several widely used baselines.

4.3 Visualization Comparison

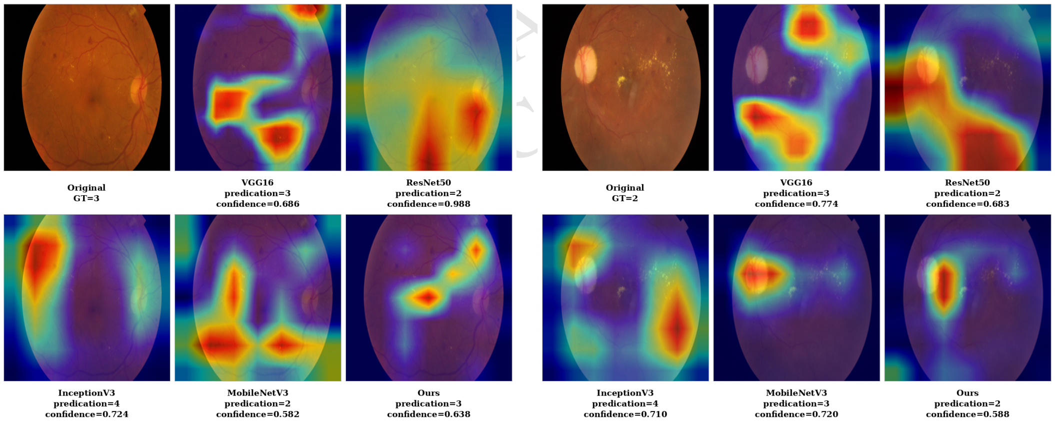

Figure 2: Visualization results from GradCAM between our method and the other baseline methods.

We further visualize class activation maps (CAMs) for representative cases to examine the spatial evidence used by different models. Fig. 2 shows two fundus images where predictions from several baselines are inconsistent with the ground truth. In the first case (true label: grade 3), VGG16 and our model correctly identify the severity, whereas ResNet50 and MobileNetV3 underestimate the grade and InceptionV3 overestimates it. The corresponding CAMs reveal that several baselines attend to scattered retinal regions, while our model concentrates on lesion-prone areas with clearer and more compact responses, suggesting improved localization of pathological cues. In the second case (true label: grade 2), multiple baselines misclassify the image as higher severity levels, whereas our model and ResNet50 predict the correct grade. The visualization indicates that misclassified models tend to highlight broader background regions, while our model focuses on localized lesion structures, leading to a more consistent grading decision. Overall, these examples suggest that our model produces more concentrated and pathology-aware activation patterns, which aligns better with clinically relevant retinal structures and contributes to more reliable DR grading.

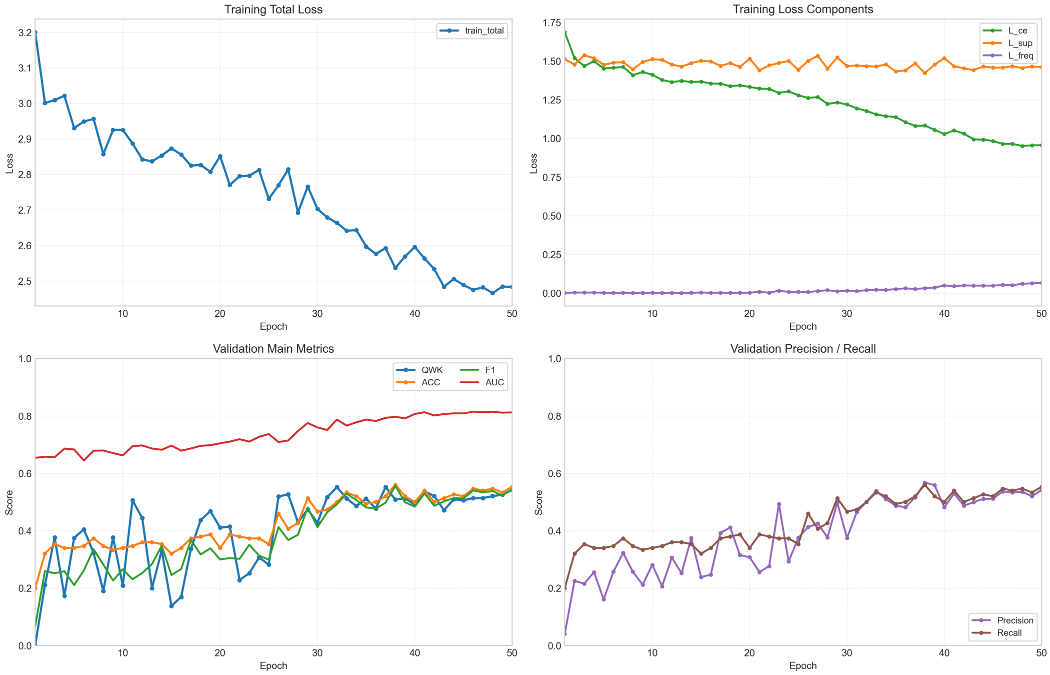

Figure 3: Learning curves over 50 epochs.

Table 2: Ablation study of the proposed PyT-SORD++ framework. Each row removes specific components from the full model.

Method	Acc ↑	QWK ↑	Prec ↑	F1 ↑	AUC ↑
PyT-SORD++ (Full)	**0.560**	0.508	**0.567**	**0.556**	**0.797**
w/o FBS	0.533	**0.572**	0.521	0.526	0.768
w/o PosBias	0.520	0.529	0.512	0.510	0.768
w/o FBS + VAB + PosBias	0.507	0.535	0.510	0.503	0.765
w/o FBS + VAB	0.500	0.514	0.477	0.485	0.752
w/o FBS + PosBias	0.487	0.454	0.482	0.478	0.740
w/o VAB	0.453	0.462	0.467	0.436	0.741
w/o VAB + PosBias	0.453	0.440	0.422	0.426	0.726

4.4 Training Dynamics and Convergence

Fig. 3 shows the training dynamics over 50 epochs. The model improves rapidly in early training, with QWK increasing from 0 to 0.51 and accuracy from 20% to 35%, while AUC rises from 0.65 to 0.69, indicating improved ranking ability. During the middle stage, the cross-entropy loss decreases steadily and validation metrics (F1, AUC, and accuracy) continue to improve with moderate fluctuations, suggesting stable optimization under the joint loss formulation. As the frequency regularization weight reaches its scheduled maximum, the model gradually incorporates frequency-domain constraints without destabilizing training. In the later stage (epochs 30–50), the training process stabilizes and the model converges. Despite occasional gradient anomalies, gradient clipping and AMP scaling maintain stable optimization. Overall, these results demonstrate consistent convergence and stable multi-objective training dynamics.

4.5 Ablation Studies

To better understand the contribution of each design component in PyT-SORD++, we conduct a series of ablation experiments by selectively removing modules from the full model. Table 2 summarizes the quantitative results. We first evaluate the contribution of each component independently. Removing the Feature Balance Strategy (FBS) reduces accuracy from 56.0% to 53.3% and decreases

AUC from 0.797 to 0.768. This result indicates that balanced integration of multi-scale representations is important for jointly capturing micro-lesion details and global anatomical context. Eliminating the Positional Bias (PosBias) also leads to a performance drop, yielding 52.0% accuracy and 0.768 AUC. The decrease suggests that spatially informed attention improves the model's ability to focus on clinically relevant retinal regions. The most substantial degradation occurs when removing the Vascular-Aware Bias (VAB). Without this component, performance drops to 45.3% accuracy and 0.741 AUC, highlighting the importance of incorporating vascular priors when modeling the interaction between local lesion cues and global anatomical structures.

When FBS and PosBias are both removed, accuracy drops further to 48.7%, with an AUC of 0.740, indicating that the pyramidal representation and spatial priors complement each other in preserving meaningful structural information. Removing both FBS and VAB results in 50.0% accuracy and 0.752 AUC, suggesting that while multi-scale feature balancing improves performance, the vascular-aware attention provides a stronger structural constraint. The configuration removing all three components (FBS, VAB, and PosBias) yields 50.7% accuracy and 0.765 AUC, which remains noticeably lower than the full model. This observation confirms that the joint presence of pyramidal feature balancing, anatomy-aware spatial priors, and vascular-guided attention contributes to the most robust representation. Finally, the variant without VAB and PosBias achieves the lowest AUC (0.726) among all configurations, further emphasizing the importance of anatomy-aware attention mechanisms.

The ablation results provide several insights into the design of PyT-SORD++. First, the pyramidal representation with feature balancing helps maintain consistent performance by preserving both micro-level lesion signals and macro anatomical context. Second, anatomy-aware attention mechanisms, particularly the vascular-aware bias, play a critical role in guiding the model toward clinically meaningful structures. Third, the combination of spatial priors and vascular-aware attention improves robustness by constraining the attention distribution to anatomically plausible regions.

5 Conclusion

We tackle reliable five-class diabetic retinopathy grading from single fundus images, aiming to detect micro-lesions in their vascular context while maintaining calibration across device and illumination shifts. We present PyT-SORD++, a single-pass pyramidal transformer that preserves fine detail and global anatomy via anatomy-aware token gating and vascular-biased micro-to-macro attention. The training objective unifies supervised classification, ordinal contrastive calibration, and Fourier-based style consistency to couple discriminative accuracy with device-tolerant behavior. On public DR benchmarks, PyT-SORD++ improves accuracy, maintains ordinal consistency, and yields better calibration and robustness than strong CNN and transformer baselines. It notably reduces large-grade misclassifications, enabling more reliable single-pass risk stratification. We will pursue device-diverse, prospective multi-center validation and integrate uncertainty-aware calibration to support clinician-in-the-loop triage and safe deployment.

6 Ethics Statement

We conducted this study using publicly available, de-identified retinal fundus datasets: APTOS 2019 retinal fundus dataset Aravind Eye Hospital and PG Institute of Ophthalmology (2019) sponsored by Aravind Eye Hospital & PG Institute of Ophthalmology (India). Ethical approval was not required as confirmed by the license attached to the open access data.

References

Aravind Eye Hospital and PG Institute of Ophthalmology. APTOS 2019 Blindness Detection. https://www.kaggle.com/competitions/aptos2019-blindness-detection, 2019. Kaggle Competition.

Dian Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3690–3698, 2020.

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

Early Treatment Diabetic Retinopathy Study Research Group. Grading diabetic retinopathy from stereoscopic color fundus photographs—an extension of the modified airlie house classification. etdrs report number 10. *Ophthalmology*, 98(5):786–806, 1991.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning*, pp. 1321–1330, 2017.

Jian Hu, Peter Schmidt, and Matthias Heinig. Microglia orchestrate retinal angiogenesis in health and diabetic retinopathy. *Angiogenesis*, 27(2):145–162, 2024.

Maximilian Ilse, Jakub M Tomczak, and Max Welling. Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712*, 2018.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yifan Tian, Phillip Isola, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020.

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*, pp. 10012–10022, 2021.

Hui Mao, Chaofei Wang, Zuxuan Wang, Caiming Xiong, Zhangyang Wang, and Zhiding Yu. Towards robust vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12042–12051, 2022.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning*, pp. 6105–6114, 2019.

Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Stein, Daniel Keysers, Jakob Uszkoreit, and Mario Lucic. Mlp-mixer: An all-mlp architecture for vision. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24261–24272, 2021.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7819–7830, 2021.

Wenhai Wang, Enze Xie, Xiang Li, Dengping Fan, Kaiming 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*, pp. 568–578, 2021.

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22–31, 2021.

Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4084–4094, 2020.

Tong Yu, Xudong Li, Yucheng Cai, Mingyuan Zhang, Shi Liu, Jiashi Li, Kuan Han, and Yunhe Wang. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5585–5594, 2022.

Wei Zhang, Jun Li, and Hao Chen. Transformer with multiple instance learning for high-resolution diabetic retinopathy grading. *IEEE Access*, 12:123456–123468, 2024.