DC-SSNET: LINEAR-TIME STATE-SPACE DIABETIC RETINOPATHY GRADING WITH THIN-STRUCTURE PRESERVATION AND OVERLAP-AWARE TRAINING

Anonymous Author(s)

ABSTRACT

Diabetic retinopathy grading from single-field fundus photographs is a five-class ordinal problem in which subtle micro-lesions and thin vessels are blurred by downscaling, global context is expensive to capture at clinical resolutions, and class imbalance with near-boundary overlap confounds intermediate grades. We introduce DC-SSNet to improve lesion visibility, encode global context efficiently, and mitigate ordinal overlap under practical computation at 512×512. The model couples lesion-aware modulation, an efficient state-space encoder, and objectives tailored to imbalance and overlap: a dynamic, cluster-aware emphasis module combines image gradients with Laplacian-of-Gaussian responses to highlight lesion-prone regions while preserving vessels; a multi-stage encoder uses selective two-dimensional state-space scans, a local–global memory split, and oriented thin-structure fusion to achieve linear-time global mixing without sacrificing fine detail; and training integrates lesion-size-aware sampling with a loss that blends a class-balanced focal term and a prototype neighborhood components analysis regularizer to reduce inter-class embedding overlap. On APTOS 2019 with a stratified 80/20 validation split, the primary metric is quadratic weighted kappa, with macro AUC and specificity secondary. Extensive experiments on a public diabetic retinopathy grading benchmark demonstrate the effectiveness of the proposed method and components. These findings show that global context can be aggregated efficiently while preserving fine structure at practical resolution and highlight directions to improve sensitivity and separation between adjacent grades for scalable DR screening.

1 INTRODUCTION

Automated analysis of retinal fundus photographs has become a central computer vision task in medical imaging, with diabetic retinopathy (DR) screening a prominent use case. DR grading from single-field fundus images is a five-class, ordinal problem that must detect minute lesions such as microaneurysms while reasoning over global retinal context. Reliable grading at clinical resolution is crucial for early referral and treatment, yet remains challenging due to variability in image quality, devices, and patient populations. This work addresses the problem of accurate, robust, and efficient DR grading from high-resolution fundus photographs.

Convolutional networks and attention-based encoders have advanced DR screening by leveraging large-scale pretraining, multi-scale features, and long-range dependencies Gulshan et al. (2016); He et al. (2016); Dosovitskiy et al. (2021); Liu et al. (2021). These advances deliver strong baselines, but practical constraints persist. Clinical-resolution inputs are frequently downsampled or tiled, which can suppress thin vessels and micro-lesions or break global context. Generic augmentations and aggressive resizing may diminish low-contrast cues, while obtaining global context at high resolution is computationally demanding. Severe class imbalance and overlap near ordinal boundaries lead to confusion among intermediate grades, and acquisition artifacts, domain shift, and view variability reduce generalization. Recent state-space models offer linear-time global mixing without quadratic attention Gu & Dao (2024); Liu et al. (2024), yet they are seldom coupled with mechanisms that preserve lesion-level detail under clinical constraints.

Addressing these limitations matters for both clinical outcomes and scalable deployment. Sensitivity to early, subtle lesions determines timely intervention, while precise separation near ordinal boundaries influences referral thresholds and resource allocation. Methods that preserve fine detail and global context can reduce missed diagnoses and false referrals. Efficiency at high resolution enables routine use on commodity hardware, and robustness to acquisition variability supports broader adoption across diverse settings.

We introduce DC-SSNet, a DR grading framework that integrates content-adaptive lesion emphasis with an efficient encoder for global and local context, and a training strategy aligned with ordinal structure and class imbalance. The core idea is to amplify lesion-prone patterns without erasing thin structures, while using attention-free sequence mixing to capture long-range relationships at a practical computational cost. The approach steers learning toward rare and borderline cases to improve separability near grade boundaries and enhance generalization across views and devices. Together, these components aim to restore clinically meaningful detail, maintain global awareness, and improve reliability under real-world variability.

Our main contributions are as follows.

- We present DC-SSNet, a high-resolution DR grading framework that preserves fine, low-contrast lesions while retaining global retinal context within a practical compute budget.
- We propose a content-adaptive emphasis mechanism that highlights lesion-prone regions and mitigates artifacts and view variability without sacrificing thin structures.
- We design a training strategy that combines targeted sampling with ordinal- and imbalance-aware supervision to improve separation near boundary grades and handle rare cases.
- We provide a systematic empirical evaluation with ablations and qualitative analyses that link model cues to clinical findings and show consistent gains on a public benchmark.

2 Related Work

2.1 CNN- and Transformer-based Diabetic Retinopathy Classification

Transfer learning with convolutional neural networks (CNNs) set strong baselines for diabetic retinopathy (DR) screening and multi-class grading from fundus photographs. Clinical-scale systems based on Inception- or ResNet-like backbones reported high sensitivity and specificity across diverse cohorts and supported the feasibility of autonomous DR screening in practice Gulshan et al. (2016); Ting et al. (2017); He et al. (2016); Tan & Le (2019). Later studies examined deeper or more parameter-efficient backbones, multi-scale preprocessing, and interpretability to stabilize performance across imaging protocols and devices; CAM-based visualizations (e.g., Grad-CAM) were used to check that highlighted evidence aligned with lesion locations Selvaraju et al. (2017). Recently, Transformer encoders have been adopted to model long-range dependencies via self-attention; ViT/DeiT and hierarchical or windowed variants such as Swin and PVT have been adapted for fundus classification through ImageNet pretraining and fine-tuning Dosovitskiy et al. (2021); Touvron et al. (2021); Liu et al. (2021); Wang et al. (2021). Deploying attention at clinical resolutions (2–6K) often requires downscaling, tiling, or multicrop pipelines that can reduce the visibility of tiny lesions and complicate training and calibration.

Compared with this line of work, we target fine-grained 5-class DR grading by preserving tiny, clustered lesions while capturing global context within a realistic compute budget. We replace quadratic self-attention with state-space layers that scale linearly and pair them with lesion-aware training and class-imbalance- and overlap-aware objectives to improve sensitivity near grade boundaries.

2.2 State Space Models for Vision and Medical Imaging

Structured state space models (SSMs) reintroduce sequence layers with very long effective context and linear-time complexity. S4 formalized stable, long-range dynamics for deep learning, and selective SSMs such as Mamba modulate input- and state-dependent updates to achieve high throughput and memory locality for long contexts Gu et al. (2022); Gu & Dao (2024). Vision adaptations take both token- and map-centric forms: Vision Mamba provides competitive visual encoders without quadratic attention, and VMamba introduces 2D selective scanning (SS2D) that aggregates global

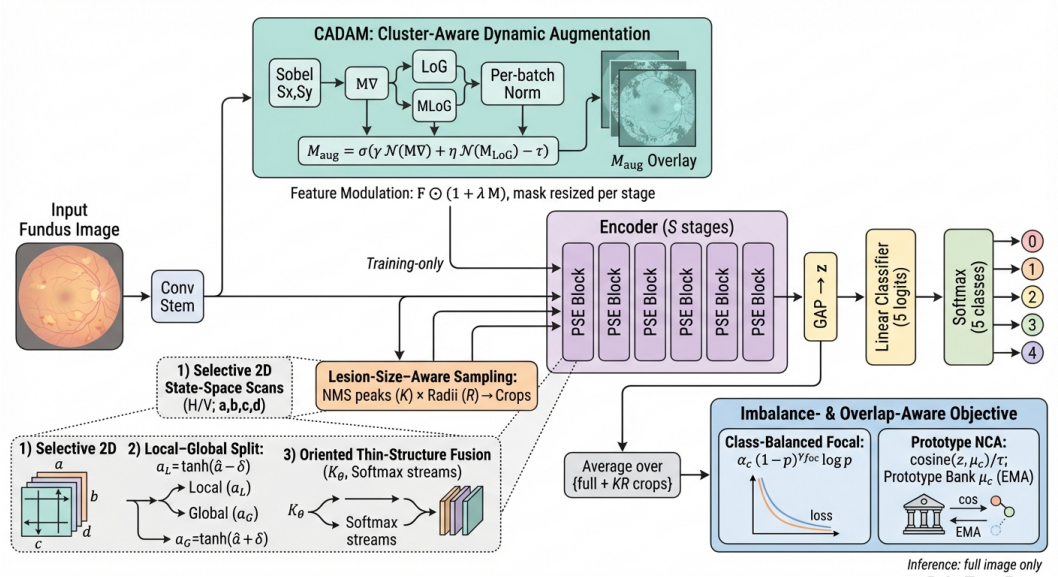

Figure 1: Overview of DC-SSNet for five-class diabetic retinopathy grading. The framework includes cluster-aware dynamic augmentation, a multi-stage encoder with progressive sensitivity encoding, lesion-size–aware sampling, and an objective that accounts for class imbalance and inter-class overlap through class-balanced focal loss and prototype regularization.

context via multidirectional recurrences over feature maps Zhu et al. (2024); Liu et al. (2024). In medical imaging, U-shaped hybrids (e.g., U-Mamba) embed SSM blocks into encoder–decoder networks to capture global context for segmentation with good accuracy–efficiency trade-offs compared with Transformer counterparts at clinical resolutions Ma et al. (2024). Despite this progress, most Mamba-based medical work focuses on dense prediction, and far fewer studies address image-level grading with subtle, high-frequency cues under weak supervision.

Building on these advances, we adapt SS2D to image-level DR severity grading by coupling efficient four-way scans with thin-structure-preserving fusion. This provides linear-time global context aggregation tailored to fundus images, enabling high-resolution inputs and improved sensitivity to micro-lesions without the overhead of full self-attention.

2.3 Multi-Scale Fusion, Targeted Augmentation, and Imbalance-/Overlap-Aware Learning

Multi-scale fusion is central to preserving small structures and global context in retinal images. Feature pyramids and U-shaped designs route high-resolution features to deeper stages, while high-resolution backbones maintain parallel streams to mitigate over-smoothing of vessels and micro-lesions Lin et al. (2017a); Ronneberger et al. (2015); Sun et al. (2019). Augmentation and normalization also affect lesion visibility: contrast-limited adaptive histogram equalization (CLAHE) can enhance low-contrast lesions, whereas mix-based policies (e.g., MixUp) may distort thin vascular structures if applied indiscriminately Zuiderveld (1994); Zhang et al. (2018). Beyond architecture and augmentation, objectives that address long-tailed distributions and near-boundary confusion have been effective: focal and class-balanced losses reweight rare or hard examples, and margin-based or contrastive formulations (e.g., LDAM, SupCon) improve separation near ambiguous grades Lin et al. (2017b); Cui et al. (2019); Cao et al. (2019); Khosla et al. (2020). In object detection, adaptive training sample selection (ATSS) provides a mechanism to emphasize dense, small targets, which translates naturally to lesion-size–aware cropping and sampling in DR classification Zhang et al. (2020).

Our pipeline combines a thin-structure-preserving U-shaped fusion pathway with efficient SS2D-based global propagation, lesion- and macula-aware augmentations that maintain vascular morphology, and class-imbalance- and overlap-aware objectives.

3 METHODOLOGY

We present DC-SSNet for five-class diabetic retinopathy (DR) grading from retinal fundus images. The model couples a cluster-aware dynamic augmentation module that emphasizes lesion-prone regions with a multi-stage encoder using progressive sensitivity encoding to capture local-to-global dependencies while preserving thin structures. Training uses an objective that accounts for class imbalance and inter-class overlap by combining a class-balanced focal term with a prototype NCA regularizer.

3.1 RETINAL DR ENCODER

Let $I \in \mathbb{R}^{B\times 3\times H_0\times W_0}$ denote a preprocessed mini-batch of RGB fundus images. A convolutional stem produces early features $F_{\text{stem}} \in \mathbb{R}^{B\times C_1\times H_1\times W_1}$. A spatial importance mask $M_{\text{aug}} \in \mathbb{R}^{B\times 1\times H_1\times W_1}$ is computed from F_{stem} by the cluster-aware dynamic augmentation module and is used to modulate features entering a multi-stage encoder. The encoder has S stages with progressive sensitivity encoding (PSE), yielding features $\{F_s\}_{s=1}^{S}$ with $F_S \in \mathbb{R}^{B\times C_S\times H_S\times W_S}$. We obtain the final representation by global average pooling $z = \mathrm{GAP}(F_S) \in \mathbb{R}^{B\times D}$ and a linear classifier that produces logits $\ell = W_{\text{cls}} z + b_{\text{cls}} \in \mathbb{R}^{B\times 5}$. Training uses the imbalance- and overlap-aware objective defined in Section 3.5.

To preserve spatial emphasis at multiple resolutions, the single mask M_{aug} computed at resolution (H_1, W_1) is bilinearly resized to each stage resolution and used to modulate stage inputs.

3.2 CLUSTER-AWARE DYNAMIC AUGMENTATION

The augmentation mask fuses gradient and Laplacian-of-Gaussian (LoG) responses computed on early features to highlight lesion clusters and thin vessels. Let S_x, S_y be fixed Sobel filters and LoG a fixed Laplacian-of-Gaussian kernel. Define channelwise gradient and LoG responses and their channel-averaged magnitudes:

$$\begin{aligned} G_x &= F_{\text{stem}} * S_x, \quad G_y = F_{\text{stem}} * S_y, \quad M_\nabla = \sqrt{\mathrm{mean}_c\big(G_x^2 + G_y^2\big)}, \\ H &= \big|F_{\text{stem}} * \mathrm{LoG}\big|, \qquad\qquad M_{\mathrm{LoG}} = \mathrm{mean}_c(H), \end{aligned} \tag{1}$$

where $*$ denotes 2D convolution applied per channel and mean_c averages over channels. A per-batch affine normalization $\mathcal{N}(M) = \frac{M-\mu(M)}{\sigma(M)+\varepsilon}$ with small $\varepsilon > 0$ is applied to M_∇ and M_{LoG}. The fused mask is

$$M_{\text{aug}} = \sigma\Big(\gamma\,\mathcal{N}(M_\nabla) + \eta\,\mathcal{N}(M_{\mathrm{LoG}}) - \tau\Big) \in \mathbb{R}^{B\times 1\times H_1\times W_1}, \tag{2}$$

where $\sigma(\cdot)$ denotes the logistic sigmoid and (γ, η, τ) are learnable scalars. For any stage-aligned feature map $F \in \mathbb{R}^{B\times C\times H\times W}$ and a mask $M \in \mathbb{R}^{B\times 1\times H\times W}$, feature modulation is defined as

$$\mathcal{M}_{\mathrm{CADAM}}(M, F) = F \odot \big(\mathbf{1} + \lambda\, M\big), \qquad \lambda > 0, \tag{3}$$

where $\odot$ denotes elementwise multiplication with broadcasting over channels.

3.3 LESION-SIZE–AWARE SAMPLING

During training, lesion-focused crops are generated to increase exposure to small, clustered lesions. Let $\mathcal{P}(M_{\text{aug}})$ return a set of K non-overlapping local maxima $\{(u_k, v_k)\}_{k=1}^{K}$ obtained by non-maximum suppression on M_{aug}. For each center (u_k, v_k) and a set of crop radii $\mathcal{R} = \{\rho_1, \ldots, \rho_R\}$, a crop is extracted from the original image I by a differentiable cropping operator $\mathrm{crop}(I, (u_k, v_k), \rho)$ followed by resizing to the network input. Each crop passes through the shared backbone to yield embeddings and logits $\{z^{(k,r)}, \ell^{(k,r)}\}$ with the same image-level label. At inference, only the full-resolution input is used.

The classification and prototype terms (Section 3.5) are computed over the union of full images and lesion-focused crops, averaged per original sample:

$$\mathcal{L}_{\text{sample}}(i) = \frac{1}{1+KR} \sum_{(k,r)\in\{0\}\cup([K]\times\mathcal{R})} \Big(\mathcal{L}_{\text{cb-focal}}(i,k,r) + \lambda_{\text{proto}}\mathcal{L}_{\text{proto}}(i,k,r)\Big), \tag{4}$$

where $(k, r) = 0$ denotes the full image, $[K] = \{1, \ldots, K\}$, and the batch loss averages $\mathcal{L}_{\text{sample}}(i)$ over $i = 1, \ldots, B$.

3.4 Progressive Sensitivity Encoding

Progressive sensitivity encoding (PSE) captures spatial dependencies using linear-complexity selective scans along the horizontal and vertical axes and separates short- and long-memory dynamics. Each PSE block combines selective 2D state-space scans, a local–global memory split, and oriented thin-structure fusion.

3.4.1 Selective 2D State-Space Scans

Given $F \in \mathbb{R}^{B\times C\times H\times W}$, per-channel parameters $(a, b, c, d) \in \mathbb{R}^{B\times C\times 1\times 1}$ are produced by a 1×1 projection, with $a = \tanh(\hat{a})$ to ensure stability. Horizontal and vertical recurrences are

$$\begin{aligned} h_{i,j} &= a \odot h_{i,j-1} + b \odot F_{i,j}, \quad Y^H_{i,j} = c \odot h_{i,j} + d \odot F_{i,j}, \quad h_{i,0} = 0, \\ v_{i,j} &= a \odot v_{i-1,j} + b \odot F_{i,j}, \quad Y^V_{i,j} = c \odot v_{i,j} + d \odot F_{i,j}, \quad v_{0,j} = 0, \end{aligned} \tag{5}$$

where (i, j) index spatial positions and $\odot$ denotes elementwise multiplication. The scan output is $Y = Y^H + Y^V$, followed by a residual connection and normalization.

3.4.2 Local–Global Memory Split

To separate short-range lesion cues from long-range context, two parallel streams share (b, c, d) and differ only in the memory coefficient via a learnable shift δ:

$$a_L = \tanh(\hat{a} - \delta), \qquad a_G = \tanh(\hat{a} + \delta), \qquad \delta \in \mathbb{R}^{B\times C\times 1\times 1}. \tag{6}$$

Applying the scan equations with a_L and a_G yields Y_L and Y_G, respectively, which favor short and long effective memory.

3.4.3 Oriented Thin-Structure Fusion

Thin retinal structures are preserved through orientation-aware fusion. Let $\{K_\theta\}_{\theta\in\Theta}$ be a small bank of oriented depthwise separable kernels. Orientation-aggregated responses for the two streams are

$$R_L = \sum_{\theta\in\Theta} \big|Y_L * K_\theta\big|, \qquad R_G = \sum_{\theta\in\Theta} \big|Y_G * K_\theta\big|, \tag{7}$$

with softmax normalization across streams to obtain attention maps $A_L, A_G \in \mathbb{R}^{B\times 1\times H\times W}$:

$$[A_L, A_G] = \text{Softmax}_{\text{streams}}([R_L, R_G]). \tag{8}$$

The fused output is

$$F_{\text{PSE}} = A_L \odot Y_L + A_G \odot Y_G, \tag{9}$$

optionally followed by a pointwise projection and residual addition.

3.5 Imbalance- and Overlap-Aware Objective

Let $z \in \mathbb{R}^{B\times D}$ denote pooled embeddings and $\ell \in \mathbb{R}^{B\times 5}$ the logits. The softmax probability for sample i and class c is $p_{i,c} = \text{Softmax}(\ell_i)_c$, and $y_i \in \{0, \ldots, 4\}$ denotes the ground-truth label. The objective combines a class-balanced focal classification term with a prototype-based NCA regularizer.

3.5.1 Class-Balanced Focal Term

Rare classes are re-weighted using effective-number-derived weights $\alpha_c > 0$, and hard examples are emphasized by the focusing parameter $\gamma_{\text{foc}} \geq 0$. The per-sample term is

$$\mathcal{L}_{\text{cb-focal}} \;=\; -\frac{1}{B}\sum_{i=1}^{B} \alpha_{y_i} \left(1 - p_{i,y_i}\right)^{\gamma_{\text{foc}}} \log p_{i,y_i}. \tag{10}$$

3.5.2 PROTOTYPE NCA REGULARIZER

To reduce inter-class embedding overlap, unit-norm class prototypes $\{\mu_c \in \mathbb{R}^D\}_{c=0}^4$ are maintained via an exponential moving average (EMA). Let $\hat{z}_i = \frac{z_i}{\|z_i\|_2}$ denote the normalized embedding and N_c^{batch} the number of class-c samples in the mini-batch. The EMA update with momentum $m \in [0, 1)$ is

$$\mu_c \leftarrow \frac{m\,\mu_c + (1-m)\,\frac{1}{\max(1, N_c^{\text{batch}})}\sum_{i:\,y_i=c}\hat{z}_i}{\left\| m\,\mu_c + (1-m)\,\frac{1}{\max(1, N_c^{\text{batch}})}\sum_{i:\,y_i=c}\hat{z}_i \right\|_2}. \tag{11}$$

With cosine similarities $s_{i,c} = \hat{z}_i^\top \mu_c$ and temperature $\tau > 0$, the NCA-style loss is

$$\mathcal{L}_{\text{proto}} \;=\; -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(s_{i,y_i}/\tau\right)}{\sum_{c \neq y_i} \exp\left(s_{i,c}/\tau\right)}. \tag{12}$$

3.5.3 TOTAL TRAINING OBJECTIVE

The overall objective combines the classification and prototype terms:

$$\mathcal{L}_{\text{total}} \;=\; \mathcal{L}_{\text{cb-focal}} \;+\; \lambda_{\text{proto}}\,\mathcal{L}_{\text{proto}}, \tag{13}$$

with trade-off $\lambda_{\text{proto}} > 0$. During training with lesion-size–aware sampling, both terms are averaged over the full image and the lesion-focused crops per sample as described above.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

4.1.1 DATASETS AND EVALUATION PROTOCOLS

We evaluate on the APTOS 2019 Blindness Detection dataset of color fundus photographs annotated with five ordinal diabetic retinopathy (DR) grades (0–4). All experiments use a stratified split of the official training set into 80% training and 20% validation to preserve class proportions. Images are resized to 512×512 pixels. During training, we use color jitter (brightness/contrast/saturation factors 0.2; hue 0.02), random horizontal/vertical flips, and normalization using ImageNet channel means and standard deviations; validation uses only resizing and normalization.

4.1.2 BASELINES

For comparison, we include published validation results on APTOS 2019 for widely used convolutional and transformer architectures. These external baselines are VGG-16, ResNet-50, Inception V3, and MobileNet, which span lightweight and deeper convolutional models reported in prior work on this benchmark. All baseline networks are trained under the same experimental protocol to ensure a fair comparison. The models are optimized using the same training schedule, data preprocessing pipeline, and evaluation criteria. The best-performing checkpoint for each method is selected based on validation performance. Following prior work in medical and ordinal classification tasks, we report multiple evaluation metrics including Quadratic Weighted Kappa (QWK), accuracy (ACC), precision, recall, F1-score, specificity (SPEC), and the area under the ROC curve (AUC). Unless otherwise specified, metrics are computed on the held-out validation split. To reduce the risk of spurious epoch-wise peaks, we load both the best score across epochs and the final score at the last epoch.

4.1.3 MODEL AND TRAINING CONFIGURATION

Our model, DC-SSNet, uses a convolutional patch embedding (7×7 stride 2 with batch normalization and GELU, followed by a 3×3 block), three stages of dual-channel state-space blocks, and a linear classification head. Each DCSS block consists of a grouped convolution with channel shuffle (4 groups), a selective two-dimensional state-space mixing module with depthwise and pointwise projections, gated mixing with learnable channel-wise parameters scaled exponentially at stage-specific

Table 1: Comparison with baseline CNN architectures. The best result for each metric is shown in bold.

Method	QWK ↑	ACC ↑	Precision ↑	Recall ↑	F1 ↑	SPEC ↑	AUC ↑
VGG16	0.738	0.673	0.667	0.673	0.665	0.918	0.860
MobileNetV3	0.622	0.587	0.576	0.587	0.577	0.897	0.826
InceptionV3	0.731	0.673	0.665	0.673	0.667	0.918	**0.900**
ResNet50	**0.777**	**0.680**	0.679	**0.680**	**0.667**	**0.920**	0.897
Ours	0.727	0.667	**0.720**	0.667	0.645	0.917	0.870

dilations $\Delta \in \{0.8, 1.0, 1.2\}$, and residual connections with batch normalization. Progressive sensitivity encoding (PSE) is enabled by default. A local stream (kernel 3, $\Delta_\ell = 0.5$) and a global stream (kernel 5, $\Delta_g = 2.0$) are fused using channel- and spatial-weighting with a learnable balance parameter. An optional content-adaptive augmentation module (CADAM) computes a Sobel gradient magnitude map on grayscale feature activations and modulates features as $\mathrm{feat} \cdot (1 + \lambda_{\mathrm{CADAM}} M_{\mathrm{aug}})$ with $\lambda_{\mathrm{CADAM}} = 0.5$ and $M_{\mathrm{aug}} \in [0, 1]$; this module is toggled in ablations. Stochastic depth with drop-path rate 0.05 is applied within selective blocks.

Training uses AdamW with learning rate 3×10^{-4}, weight decay 0.05, and parameter grouping to exclude biases and normalization parameters from weight decay. We train for 100 epochs with automatic mixed precision, gradient clipping at a global norm of 5.0, and batch size 8 for the main comparison and 4 for ablations. The loss combines class-balanced focal loss (effective-number weighting with $\beta = 0.9999$) and a prototype NCA regularizer ($\lambda_{\mathrm{proto}} = 0.1$ unless ablated). Prototypes are kept as exponential moving averages of per-class feature means and are used within the NCA-style objective to reduce inter-class embedding overlap. In our implementation, when computed on detached features, the prototype regularizer does not backpropagate into the backbone and serves as a regularizer for the prototype space.

4.2 Main Performance Comparison

Table 1 reports the quantitative comparison between the proposed DC-SSNet and the baseline models. Among all compared methods, ResNet50 achieves the best overall classification accuracy (68.0%) and the highest QWK score (0.777), suggesting that deeper residual architectures remain competitive for this task. However, our proposed approach demonstrates strong performance across several complementary metrics. Specifically, our method achieves an AUC of 0.870, outperforming VGG16 (0.860) and MobileNetV3 (0.826), while remaining competitive with ResNet50 (0.897). In terms of precision, our model reaches 0.720, which is the highest among all evaluated methods, indicating that the proposed approach produces more reliable positive predictions. Compared with lightweight architectures such as MobileNetV3, our approach improves the QWK score by +0.105 and accuracy by +8.0 %, demonstrating substantially stronger ordinal prediction consistency. Furthermore, our method maintains high specificity (0.917), comparable to other deep CNN baselines. Although ResNet50 slightly outperforms our method in terms of accuracy and QWK, the proposed model provides a more balanced trade-off across precision, specificity, and AUC. This suggests that the proposed design improves prediction reliability while maintaining competitive overall classification performance.

4.3 Training Dynamics and Learning Curves

Figure 2 illustrates the training dynamics over 100 epochs, where evaluation metrics are plotted using results sampled every five epochs for clarity. As training progresses, the optimization exhibits a steady decrease in the training loss, dropping from 5.03 at the first epoch to 1.99 at epoch 100, indicating stable optimization and effective parameter updates. Correspondingly, performance metrics show a consistent upward trend despite moderate fluctuations during the early training phase. In particular, the quadratic weighted kappa (QWK) improves substantially from 0.30 to 0.80, reflecting progressively better alignment with the ordinal grading structure. Similarly, accuracy increases from 34.7% to 64.7%, while F1-score rises from 0.29 to 0.64, suggesting balanced improvements in both precision and recall.

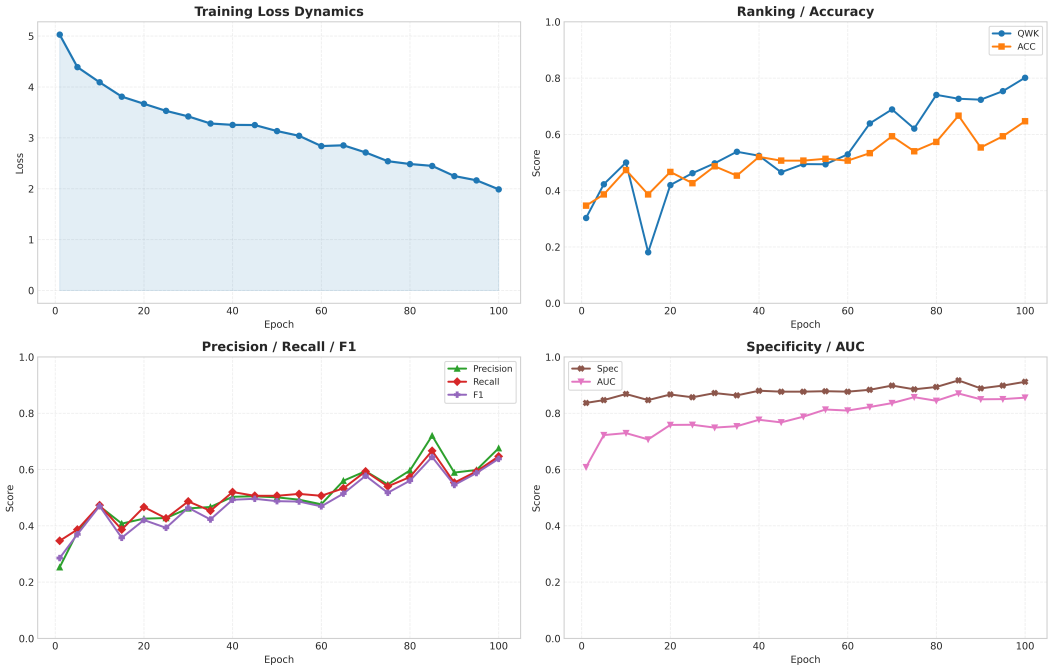

Figure 2: Learning curves on APTOS 2019.

Table 2: Ablation study of DC-SSNet. Each variant removes a specific component from the full model.

Method	QWK ↑	ACC ↑	Precision ↑	Recall ↑	F1 ↑	SPEC ↑	AUC ↑
w/o CADAM	0.639	0.600	0.590	0.600	0.581	0.900	0.843
w/o PSE	**0.757**	0.653	0.641	0.653	0.643	0.913	**0.882**
w/o Overlap	0.610	0.580	0.630	0.580	0.557	0.895	0.831
DC-SSNet (Ours)	0.727	**0.667**	**0.720**	**0.667**	**0.645**	**0.917**	0.870

Notably, most metrics begin to stabilize after approximately 60 epochs, where the model enters a relatively stable convergence regime. During this stage, performance continues to improve gradually, with AUC increasing from 0.61 in the early stage to 0.86 by the end of training, and specificity consistently remaining above 0.88 in the later epochs. The overall trend demonstrates that the proposed training strategy yields stable convergence while progressively enhancing both discriminative ability and ordinal consistency. These observations confirm the robustness of the optimization process and the effectiveness of the proposed framework in learning reliable representations for diabetic retinopathy grading.

4.4 Ablation Studies

To analyze the contribution of each component in DC-SSNet, we conduct ablation experiments by removing individual modules while keeping the rest of the architecture and training configuration unchanged. The evaluated components correspond to the key design elements introduced in the method section, including the content-adaptive emphasis mechanism (CADAM), the patch structure enhancement module (PSE), and the overlapping patch embedding strategy. All variants are trained under the same settings, and the best-performing checkpoints are used for evaluation.

4.4.1 Effect of CADAM

We first evaluate the effect of removing the content-adaptive emphasis mechanism (CADAM). As shown in Table 2, removing CADAM leads to a clear degradation across all major metrics. The QWK score drops from 0.727 to 0.639, while classification accuracy decreases from 66.7% to

60.0%. Similarly, the AUC decreases from 0.870 to 0.843, and the F1 score falls from 0.645 to 0.581. These results indicate that CADAM plays an important role in highlighting lesion-prone regions and guiding the network toward clinically meaningful cues, which improves ordinal consistency and classification reliability.

4.4.2 Effect of PSE

Next, we remove the patch structure enhancement (PSE) module to assess its impact. Without PSE, the model achieves a QWK score of 0.757 and an AUC of 0.882, while the overall accuracy slightly decreases to 65.3% compared with 66.7% for the full model. The precision and F1 score also decline to 0.641 and 0.643, respectively. These observations suggest that PSE contributes to more stable feature representation by strengthening local structural information, which benefits overall classification consistency.

4.4.3 Effect of the Overlap Penalty

Finally, we evaluate the overlapping patch embedding strategy. When the overlap mechanism is removed, the model performance drops substantially. The QWK score decreases to 0.610, and accuracy declines to 58.0%, representing the largest degradation among all ablation settings. The AUC also drops from 0.870 to 0.831, while the F1 score decreases to 0.557. This confirms that overlapping patches are critical for preserving spatial continuity and capturing fine-grained lesion patterns across patch boundaries.

Overall, these ablation results demonstrate that each component contributes to the effectiveness of DC-SSNet. In particular, the content-adaptive emphasis mechanism and overlapping patch embedding play key roles in improving ordinal prediction quality and maintaining sensitivity to subtle retinal lesions.

5 Conclusion

Grading diabetic retinopathy from single-field fundus photographs is ordinal and class-imbalanced, demanding sensitivity to tiny lesions while preserving global context under practical compute. We introduce DC-SSNet, a high-resolution framework that couples content-adaptive emphasis with linear-time state-space encoding to fuse local detail and global context. Training addresses imbalance and label overlap via class-balanced focal loss, prototype regularization, and lesion-size-aware sampling. On a public benchmark, DC-SSNet attains competitive ordinal agreement with modest computation. Ablations reveal that edge-centric emphasis and local-global fusion can misallocate attention and that an overlap penalty does not sharpen boundaries, highlighting confusion between adjacent grades. These findings support interpretable, compute-efficient retinal models while underscoring sensitivity to cohort shift and fairness. We will replace edge-centric emphasis with lesion-aware stage- and channel-gated modulation, tie prototypes to gradients, and validate at higher resolution across multi-center cohorts to improve separation of adjacent grades.

6 Ethics Statement

This research study used only the publicly data made accessible in open access by the APTOS 2019 Blindness Detection competition Aravind Eye Hospital and PG Institute of Ophthalmology (2019) on Kaggle, 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.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1567–1578, 2019.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9268–9277, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2024.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 8086–8099, 2022.

Varun Gulshan, Lily Peng, Marc Coram, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22): 2402–2410, 2016. doi: 10.1001/jama.2016.17216.

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 (CVPR)*, pp. 770–778, 2016.

Prannay Khosla, Piotr Teterwak, Chen Wang, et al. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 18661–18673, 2020.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, et al. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2117–2125, 2017a.

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 (ICCV)*, pp. 2980–2988, 2017b.

Yutong Liu, Zengqiang Chen, Haoning Xu, et al. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.00939*, 2024.

Ze Liu, Yutong Lin, Yue Cao, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021.

Jun Ma, Lequan Yu, and Jing Wang. U-mamba: Enhancing u-net with state space model for medical image segmentation. *arXiv preprint arXiv:2401.06954*, 2024.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241. Springer, 2015.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5693–5703, 2019.

Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 6105–6114, 2019.

Daniel Shu Wei Ting, Carol Yim-Lui Cheung, Gilbert Lim, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*, 318(22):2211–2223, 2017. doi: 10.1001/jama.2017.18152.

Hugo Touvron, Matthieu Cord, Matthijs Douze, et al. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 10347–10357, 2021.

Wenhai Wang, Enze Xie, Xiang Li, et al. Pyramid vision transformer: A versatile backbone for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 568–578, 2021.

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

Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9759–9768, 2020.

Qi Zhu, Zicheng Li, Wengang Zhou, et al. Vision mamba: Efficient visual representation learning with state space model. *arXiv preprint arXiv:2401.04081*, 2024.

Karel Zuiderveld. Contrast limited adaptive histogram equalization. In *Graphics Gems IV*, pp. 474–485. Academic Press, 1994.