CROSS-SCALE CHANNEL ATTENTION WITH ORDINAL-CATEGORICAL DUAL HEADS AND UNCERTAINTY-GATED SELF-TRAINING FOR DIABETIC RETINOPATHY GRADING

Anonymous Author(s)

ABSTRACT

Diabetic retinopathy (DR) grading from fundus photographs demands sensitivity to small, sparse lesions, respect for the five-level ordinal scale, and robustness to long-tailed and noisy labels, while interpretability is desirable despite scarce lesion annotations. We propose a fundus-only grader that couples a Cross-Scale Channel Attention (CSCA) module with dual heads—a categorical softmax head and a CORAL ordinal head—trained end-to-end under a hybrid objective. The categorical head uses class-balanced focal loss to handle imbalance, and the ordinal head uses a focal-augmented ordinal loss to enforce monotonicity and reduce severe ordinal errors. By integrating cross-scale attention, ordinal modeling, and noise-aware learning without lesion-level supervision, the approach advances DR grading under realistic constraints. To mitigate label noise, we refine pseudo-labels via mutual-information gating with Monte Carlo dropout and apply a lightweight prediction-consistency regularizer across two augmentations for accepted pseudo-labels. Evaluated on APTOS 2019, the proposed method achieves a quadratic weighted kappa of 0.85, an accuracy of 73.33%, a macro-F1 score of 73.23%, and a macro one-vs-rest AUC of 91.54%.

1 INTRODUCTION

Automated analysis of retinal fundus photographs is a long-standing goal in computer vision for healthcare, with diabetic retinopathy (DR) grading serving as a prominent benchmark and clinically impactful task. DR severity is assigned on a five-level ordered scale, and early detection hinges on recognizing tiny, sparse lesions while avoiding large out-of-order errors that contradict disease progression Group (1987). In routine screening, datasets are often imbalanced, severe cases are rare, and labels near grade boundaries can be noisy. We address the problem of learning a robust, interpretable DR grader that is simultaneously sensitive to microlesions, consistent with the ordinal scale, and resilient to long-tailed and imperfect labels.

Deep learning has substantially improved DR detection and grading Gulshan et al. (2016); Ting et al. (2017), and advances in attention and multi-scale modeling have boosted lesion saliency Hu et al. (2018); Woo et al. (2018); Wang et al. (2020b); Li et al. (2019). Ordinal approaches explicitly encode label ordering Cao et al. (2020), and uncertainty-aware strategies seek robustness to label noise and distributional shift Gal & Ghahramani (2016); Xie et al. (2020). However, gaps remain. Channel attention is typically applied within a single scale, while multi-branch pyramids improve context at the cost of complexity and reduced interpretability at readout. Treating grades as nominal improves separability but can increase non-adjacent mistakes; enforcing ordering alone can dilute minority-class discrimination. Confidence-only pseudo-labeling risks confirmation bias and may disproportionately exclude rare classes. Finally, many pipelines trade off clinical interpretability when departing from a global-average-pooling (GAP) readout that supports CAM/Grad-CAM explanations.

These limitations matter in practice. Screening systems must highlight subtle, multi-scale cues such as microaneurysms and hemorrhages, yet remain faithful to the disease continuum to avoid clinically implausible jumps in predicted severity. They must also handle long-tailed distributions and imperfect labels without relying on external data or heavy architectural overhead, enabling reproducible

evaluation and deployment on common hardware. Maintaining compatibility with saliency-based explanations is important for clinician trust and for auditing model behavior in safety-critical settings.

We propose a single-stream framework that couples cross-scale feature selection, ordinal-aware supervision, and uncertainty-guided label refinement. First, we introduce Cross-Scale Channel Attention (CSCA), a descriptor-level mechanism that aggregates and re-weights channel descriptors across backbone stages. This design injects multi-scale context without spatial attention maps or multi-branch towers, and preserves a GAP-based readout for compatibility with CAM/Grad-CAM. Second, we adopt dual-head hybrid supervision: a categorical softmax head maintains class separation beneficial for minority classes and referral decisions, while an ordinal head (e.g., CORAL Cao et al. (2020)) enforces monotonic thresholds to reduce out-of-order errors. Third, we refine noisy labels via uncertainty-guided self-training that accepts pseudo-labels only when predictions are confident and exhibit low epistemic uncertainty, measured by mutual information from stochastic predictions; class-aware gating mitigates over-pruning of rare classes. We evaluate on APTOS 2019 with image-level labels only, using no external datasets or annotations, and we report internal validation under this protocol without claims of cross-dataset generalization.

Our main contributions are as follows.

- We introduce Cross-Scale Channel Attention (CSCA), a single-stream, descriptor-level cross-scale attention module that enhances microlesion saliency while retaining a GAP-based, CAM-compatible readout.
- We design a dual-head hybrid supervision scheme that combines categorical and ordinal predictions to jointly promote minority-class separability and adherence to the ordered severity scale.
- We develop an uncertainty-guided self-training procedure that refines noisy labels using confidence and mutual-information gating without external unlabeled data.
- On APTOS 2019, our approach improves DR grading under internal validation, yielding gains in quadratic weighted kappa, accuracy, macro-F1, and macro AUC, with claims limited to this dataset and evaluation regime.

2 Related Work

2.1 Backbones, Attention, Multi-Scale Fusion, and Interpretability for Fundus Classification

For fundus-based DR screening, models must capture subtle, small lesions while retaining global context and remaining compatible with GAP-driven interpretability. Early systems showed that convolutional neural networks (CNNs) with global average pooling (GAP) perform well across diverse cohorts and offer a practical path for clinical deployment Szegedy et al. (2015); Gulshan et al. (2016); Ting et al. (2017). Subsequent architectural advances improved capacity and trainability: residual learning helped address vanishing gradients and enabled much deeper models He et al. (2016), while dense connectivity promoted feature reuse and multi-scale propagation with good parameter efficiency Huang et al. (2017). Lightweight channel attention further improved the accuracy–efficiency trade-off; widely used designs include squeeze-and-excitation (SE), CBAM, efficient channel attention (ECA), selective kernel fusion, and frequency channel attention Hu et al. (2018); Woo et al. (2018); Wang et al. (2020b); Li et al. (2019). To better preserve fine details while adding broader context, multi-scale aggregation has been studied extensively. Representative designs include top–down pyramids (FPN) and multi-branch high-resolution networks (HRNet), both adopted in medical imaging for lesion-centric analysis Lin et al. (2017). A related line of work models global context via non-local operations and similar modules, such as Non-local Networks, GCNet, Gather-Excite Hu et al. (2018), and GloRe. Recently, transformer-based hierarchical backbones (e.g., Swin Transformer, PVT) have provided multi-scale token representations with global receptive fields and have been adopted widely in vision.

Weakly supervised localization via Class Activation Mapping (CAM) has been widely used to produce interpretable heatmaps from GAP-based classifiers Zhou et al. (2016), with variants that broaden coverage and improve spatial fidelity Selvaraju et al. (2017); Chattopadhay et al. (2018);

Wang et al. (2020a); Jiang et al. (2021). These techniques suit fundus imagery, where clinically meaningful cues include small, scattered red lesions and vessel-adjacent abnormalities, but saliency reliability depends on backbone choice, feature resolution, and method sensitivity, and thus requires sanity checks before drawing clinical conclusions Adebayo et al. (2018). Our approach adopts a single-stream Cross-Scale Channel Attention (CSCA) module, which preserves a single-stream pathway and a GAP-based readout, maintaining compatibility with CAM/Grad-CAM while adding cross-level context that helps preserve microlesions at the final readout resolution.

2.2 Ordinal-Aware Learning and Imbalance-Aware Optimization

DR grading follows an ordered severity scale formalized by ETDRS Group (1987), motivating objectives that respect ordinal structure. Beyond classical ordinal formulations Niu et al. (2016); Rothe et al. (2015), recent deep ordinal methods include CORAL, which shares a rank-consistent weight vector across cumulative thresholds Cao et al. (2020), CORN, which conditions binary sub-tasks on preceding outcomes, and DORN, which discretizes continuous targets and optimizes ordinal relations via classification with learned intervals. In medical imaging, ordinal objectives are used to better align predictions with disease progression and reduce non-adjacent errors relative to nominal softmax training. In parallel, long-tailed class distributions complicate DR training; focal modulation and effective-number reweighting emphasize minority and hard examples without excessive overfitting Lin et al. (2017); Cui et al. (2019). Multi-task and dual-head formulations that combine nominal classification with ordinal or regression supervision have also been explored in age estimation and medical grading, aiming to balance categorical separability with progression-aware ordering. Our approach couples a class-balanced focal loss (categorical head) with a focal-augmented CORAL loss (ordinal head) using fixed loss weights. This maintains categorical separation important for minority classes while CORAL-based monotonic thresholding discourages non-adjacent errors under long-tailed distributions.

2.3 Semi-/Self-Supervised Learning and Uncertainty-Guided Pseudo-Labeling

Semi- and self-supervised methods improve label efficiency through consistency regularization and pseudo-labeling. Mean Teacher enforces student–teacher agreement under perturbations; FixMatch combines weak–strong augmentation consistency with confidence-thresholded pseudo-labels; and Unsupervised Data Augmentation (UDA) uses weak–strong consistency with advanced augmentations Xie et al. (2020). Prototype-based assignments (e.g., PAWS) further improve sample efficiency. Reliable uncertainty is critical for selective training and robust deployment. Monte Carlo Dropout and deep ensembles provide strong baselines for epistemic uncertainty Gal & Ghahramani (2016); Lakshminarayanan et al. (2017), and post hoc calibration such as temperature scaling can improve probability alignment Guo et al. (2017). Mutual-information criteria computed over stochastic predictions provide an uncertainty signal complementary to confidence and entropy and have been used to gate pseudo-label acceptance under noisy labels. Our approach uses uncertainty-guided self-training label refinement within the labeled pool by incorporating mutual information alongside confidence, applying class-aware thresholds, and using uncertainty-dependent sharpening and weighting to mitigate confirmation bias under long-tailed DR distributions.

3 Methodology

DR grading from color fundus images requires (i) sensitivity to small, sparse microlesions across scales, (ii) respect for the ordinal structure of disease severity under long-tailed class distributions, and (iii) robustness to label noise typical of image-level annotations. We address these needs with a single-stream, end-to-end framework (as illustrated in Figure 1) that pairs a Cross-Scale Channel Attention (CSCA) module for multi-scale evidence selection with dual prediction heads, a categorical softmax head and a CORAL ordinal head, trained with a hybrid objective tuned to imbalance and ordering. To limit the impact of label noise, we refine training labels using uncertainty-gated pseudo-labels derived from mutual information (MI) estimated via Monte Carlo (MC) dropout, and optionally add a prediction-consistency regularizer. The ordinal head is used only during training; all reported metrics (QWK, accuracy, macro-F1, macro one-vs-rest AUC) and all inference rely on the categorical head. For interpretability, CAM/Grad-CAM are computed over the fused CSCA feature.

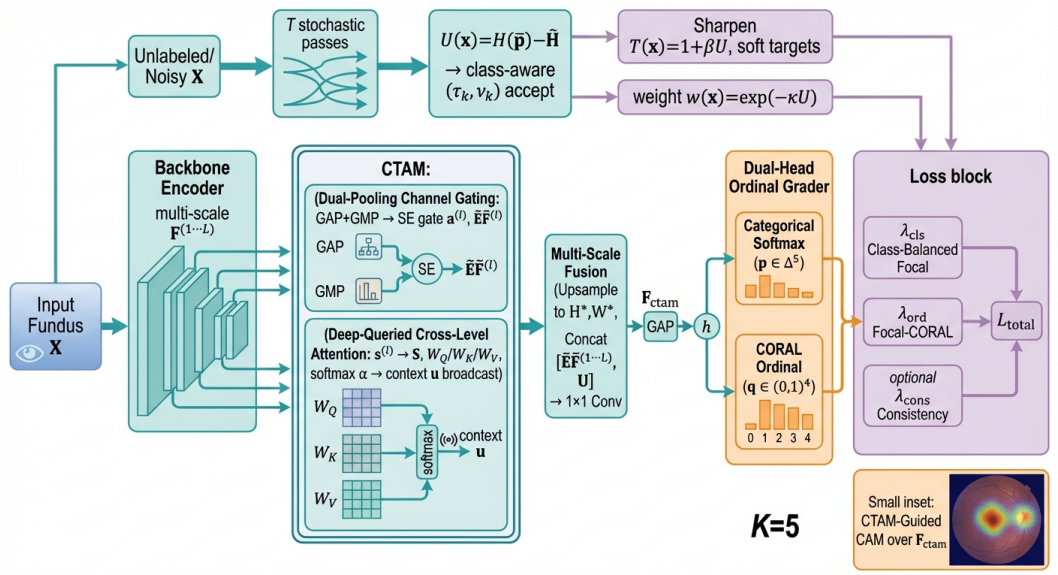

Figure 1: Architecture of the Proposed Framework. Illustrating the integration of Cross-Scale Channel Attention (CSCA), Dual-Head Grader (Categorical & CORAL), and Uncertainty-Guided Pseudo-Labeling.

Notation and symbols used once and then consistently are as follows. We set $K = 5$ for the number of grades and $L = 4$ for the backbone scales (DenseNet-121 blocks). The input batch of color fundus images is $X \in \mathbb{R}^{B\times 3\times H\times W}$, where B is the batch size. The feature map at scale l is $F^{(l)} \in \mathbb{R}^{B\times C_l\times H_l\times W_l}$, and (H_L, W_L) denotes the deepest spatial resolution. The shared descriptor dimension is d, the injected context width is C_u, and the fused channel dimension is C_f. GAP and GMP indicate global average and max pooling over spatial dimensions, $[\cdot;\cdot]$ denotes channel-wise concatenation, and $\mathrm{Broadcast}(v, H, W)$ tiles $v \in \mathbb{R}^{B\times C}$ to $\mathbb{R}^{B\times C\times H\times W}$.

3.1 Framework Overview

We use CSCA to capture cross-scale evidence for subtle lesions in a pathway compatible with global pooling, a dual-head grader to model class ordering alongside imbalance-aware categorical learning, and MI-gated pseudo-labeling to reduce the effect of noisy labels. Given X, a DenseNet-121 backbone produces a hierarchy of feature maps $\{F^{(l)}\}_{l=1}^{L}$, $F^{(l)} \in \mathbb{R}^{B\times C_l\times H_l\times W_l}$, with $L = 4$ and $C_l \in \{256, 512, 1024, 1024\}$ at the outputs of the four dense blocks. The CSCA module applies per-scale dual-pooling channel gating and deep-queried cross-level descriptor attention, then upsamples and fuses all scales at the deepest resolution via a single 1×1 convolution to produce $F_{\mathrm{csca}} \in \mathbb{R}^{B\times C_f\times H_L\times W_L}$. Global average pooling yields $h \in \mathbb{R}^{B\times C_f}$ for two heads: (i) a K-way categorical softmax head p and (ii) a CORAL ordinal head q. Training uses a hybrid objective with class-balanced focal loss on the categorical head and focal-augmented ordinal loss on the CORAL head. Uncertainty-gated pseudo-labels (from the categorical head) and an optional consistency penalty are used in a second training stage. Inference and all metrics use the categorical head only.

3.2 Cross-Scale Channel Attention (CSCA)

3.2.1 Dual-Pooling Channel Gating (Per Scale)

To elevate lesion-relevant channels at each resolution, we apply a squeeze and excitation style gate with dual pooled descriptors. For each l,

$$
\begin{aligned}
\mathbf{g}_{\text{avg}}^{(l)} &= \text{GAP}\Big(F^{(l)}\Big) \in \mathbb{R}^{B\times C_l\times 1\times 1}, \quad \mathbf{g}_{\max}^{(l)} = \text{GMP}\Big(F^{(l)}\Big) \in \mathbb{R}^{B\times C_l\times 1\times 1}, \\
\mathbf{g}^{(l)} &= \Big[\mathbf{g}_{\text{avg}}^{(l)}; \mathbf{g}_{\max}^{(l)}\Big] \in \mathbb{R}^{B\times (2C_l)\times 1\times 1}, \\
\mathbf{u}^{(l)} &= \phi\big(W_1^{(l)}\mathbf{g}^{(l)}\big) \in \mathbb{R}^{B\times (C_l/r)\times 1\times 1}, \quad \mathbf{a}^{(l)} = \sigma\big(W_2^{(l)}\mathbf{u}^{(l)}\big) \in \mathbb{R}^{B\times C_l\times 1\times 1}, \\
\tilde{F}^{(l)} &= F^{(l)} \odot \mathbf{a}^{(l)} \in \mathbb{R}^{B\times C_l\times H_l\times W_l},
\end{aligned}
\tag{1}
$$

with $r = 4$, $W_1^{(l)} \in \mathbb{R}^{(C_l/r)\times(2C_l)\times 1\times 1}$, $W_2^{(l)} \in \mathbb{R}^{C_l\times(C_l/r)\times 1\times 1}$, $\phi = \text{ReLU}$, $\sigma = \text{sigmoid}$, and $\odot$ is broadcasted channel-wise multiplication.

3.2.2 Deep-Queried Cross-Level Descriptor Attention

To consolidate cross-scale evidence while preserving a single-stream readout, we summarize each gated scale by GAP and attend across levels using the deepest scale as the query. Define $\mathbf{s}^{(l)} = \text{GAP}\Big(\tilde{F}^{(l)}\Big) \in \mathbb{R}^{B\times C_l}$. Project to a shared dimension $d = 256$:

$$
\begin{aligned}
\mathbf{q} &= \mathbf{s}^{(L)}W_Q^{(L)} \in \mathbb{R}^{B\times d}, \quad W_Q^{(L)} \in \mathbb{R}^{C_L\times d}, \\
\mathbf{k}^{(l)} &= \mathbf{s}^{(l)}W_K^{(l)} \in \mathbb{R}^{B\times d}, \quad W_K^{(l)} \in \mathbb{R}^{C_l\times d}, \\
\mathbf{v}^{(l)} &= \mathbf{s}^{(l)}W_V^{(l)} \in \mathbb{R}^{B\times d}, \quad W_V^{(l)} \in \mathbb{R}^{C_l\times d}.
\end{aligned}
\tag{2}
$$

Let $K = [\mathbf{k}^{(1)}; \ldots; \mathbf{k}^{(L)}] \in \mathbb{R}^{B\times L\times d}$, $V = [\mathbf{v}^{(1)}; \ldots; \mathbf{v}^{(L)}] \in \mathbb{R}^{B\times L\times d}$. Attention over levels:

$$
\boldsymbol{\alpha} = \text{softmax}\left(\frac{\mathbf{q}\,K^\top}{\sqrt{d}}\right) \in \mathbb{R}^{B\times L}, \quad \mathbf{c} = \sum_{l=1}^{L} \boldsymbol{\alpha}_{(:,l)} \odot \mathbf{v}^{(l)} \in \mathbb{R}^{B\times d}. \tag{3}
$$

Project and broadcast to the deepest grid:

$$
\mathbf{u} = \mathbf{c}W_C \in \mathbb{R}^{B\times C_u}, \quad W_C \in \mathbb{R}^{d\times C_u}, \quad U = \text{Broadcast}(\mathbf{u}, H_L, W_L) \in \mathbb{R}^{B\times C_u\times H_L\times W_L}, \tag{4}
$$

with $C_u = 64$.

3.2.3 Multi-Scale Upsampling, Fusion, and Readout

We then align spatial resolutions and fuse cross-scale information in a single, CAM-compatible stage. We upsample every gated map to the deepest resolution and fuse once with a single 1×1 convolution:

$$
\begin{aligned}
\hat{F}^{(l)} &= \text{Upsample}\Big(\tilde{F}^{(l)} \to (H_L, W_L)\Big) \in \mathbb{R}^{B\times C_l\times H_L\times W_L}, \quad l = 1, \ldots, L, \\
F_{\text{csca}} &= \text{Conv}_{1\times 1}\Big([\hat{F}^{(1)}; \hat{F}^{(2)}; \hat{F}^{(3)}; \hat{F}^{(4)}; U] \to C_f\Big) \in \mathbb{R}^{B\times C_f\times H_L\times W_L}, \\
h &= \text{GAP}(F_{\text{csca}}) \in \mathbb{R}^{B\times C_f},
\end{aligned}
\tag{5}
$$

where $C_f = C_L = 1024$ to preserve classifier capacity and CAM compatibility. CAM/Grad-CAM are computed over F_{csca} using the categorical head weights.

3.3 Dual-Head Grader: Categorical and CORAL Heads

3.3.1 Categorical Softmax Head

The categorical head maps h to logits z and probabilities p:

$$
z = W_c h + \mathbf{b}_c \in \mathbb{R}^{B\times K}, \quad p_{i,k} = \frac{\exp(z_{i,k})}{\sum_{j=1}^{K}\exp(z_{i,j})}, \quad i = 1, \ldots, B,\ k = 1, \ldots, K. \tag{6}
$$

All metrics and predictions use p only.

3.3.2 CORAL ORDINAL HEAD WITH MONOTONE BIAS ENFORCEMENT

We adopt CORAL with a shared weight vector and ordered biases. Let $w_{\mathrm{ord}} \in \mathbb{R}^{C_f}$ and $\mathbf{b}_{\mathrm{ord}} \in \mathbb{R}^{K-1}$ with $b_{\mathrm{ord},1} \le \cdots \le b_{\mathrm{ord},K-1}$. Threshold logits and cumulative probabilities are

$$g_{i,k} = w_{\mathrm{ord}}^{\top} h_i - b_{\mathrm{ord},k}, \quad q_{i,k} = \sigma(g_{i,k}), \quad k = 1, \ldots, K-1. \tag{7}$$

To guarantee b_{ord} is non-decreasing, we parameterize it as a cumulative sum of non-negative increments:

$$\delta_k = \mathrm{softplus}(\theta_k) \ge 0, \quad b_{\mathrm{ord},k} = \sum_{j=1}^{k} \delta_j, \quad \theta_k \in \mathbb{R}. \tag{8}$$

Targets are cumulative indicators $t_{i,k} = \mathbb{I}[y_i \ge k]$. We apply focal modulation to the ordinal subtasks:

$$\mathcal{L}_{\mathrm{ord}} = \frac{1}{|D_{\mathrm{sup}}|} \sum_{i \in D_{\mathrm{sup}}} \sum_{k=1}^{K-1} \Big(t_{i,k} \, (1-q_{i,k})^{\gamma_{\mathrm{ord}}} \, (-\log q_{i,k}) + (1-t_{i,k}) \, q_{i,k}^{\gamma_{\mathrm{ord}}} \, (-\log(1-q_{i,k})) \Big), \tag{9}$$

with $\gamma_{\mathrm{ord}} = 2$. By default, $\mathcal{L}_{\mathrm{ord}}$ is computed on D_{sup} only (we do not apply ordinal loss to pseudo-labeled samples).

3.4 UNCERTAINTY-GUIDED PSEUDO-LABEL REFINEMENT

To reduce the influence of noisy image-level labels, we refine training supervision using MI-gated pseudo-labels from the categorical head. MC dropout is used exclusively to estimate MI; it is disabled during both training and standard inference, ensuring that the deployed model remains deterministic.

3.4.1 MC DROPOUT PLACEMENT AND MI ESTIMATION

DenseNet-121 does not include dropout by default. For MI estimation only, we insert dropout layers immediately upstream of CSCA at each tapped scale (i.e., on the tensors $\{F^{(l)}\}$ prior to channel gating) and one at the fused readout:

$$F^{(l)} \xrightarrow{\mathrm{Dropout}(p_l)} F^{(l)}, \quad l = 1, \ldots, 4; \qquad h \xrightarrow{\mathrm{Dropout}(p_h)} h, \tag{10}$$

with $(p_1, p_2, p_3, p_4, p_h) = (0.10, 0.10, 0.20, 0.20, 0.20)$. These dropout layers are active only during MI estimation; they are disabled during both training and standard inference. During MI estimation, batch normalization is set to eval mode to fix running statistics. For a training image x, we run $T = 10$ stochastic forward passes to obtain $\{p^{(t)}(x)\}_{t=1}^{T}$ from the categorical head and compute

$$\begin{aligned} &\bar{p}(x) = \tfrac{1}{T} \sum_{t=1}^{T} p^{(t)}(x), \quad H(\bar{p}) = -\sum_{k=1}^{K} \bar{p}_k \log \bar{p}_k, \quad \bar{H} = \tfrac{1}{T} \sum_{t=1}^{T} \Big(-\sum_{k=1}^{K} p_k^{(t)} \log p_k^{(t)} \Big), \\ &\mathrm{MI}(x) = H(\bar{p}) - \bar{H}. \end{aligned} \tag{11}$$

3.4.2 GATING THRESHOLDS, TEMPERATURE SHARPENING, AND TARGETS

Let $\hat{y}(x) = \arg\max_k \bar{p}_k(x)$. We accept a pseudo-label if $\max_k \bar{p}_k(x) \ge \tau \quad \text{and} \quad \mathrm{MI}(x) \le \nu_{\hat{y}(x)}$, with a global confidence threshold $\tau = 0.9$ and class-wise MI cutoffs $\{\nu_c\}_{c=0}^{K-1}$ set to the 30th percentile (quantile $q = 0.3$) of the per-class MI distribution measured on the training fold under $T = 10$ MC passes (per predicted class). For accepted samples, we form a soft target by temperature sharpening of the mean probabilities:

$$T(x) = T_{\min} + (T_{\max} - T_{\min}) \cdot \mathrm{clip}\left(\frac{\mathrm{MI}(x)}{\nu_{\hat{y}(x)}}, \, 0, \, 1 \right), \quad \tilde{p}(x) = \mathrm{Softmax}\left(\frac{\log \bar{p}(x)}{T(x)} \right), \tag{12}$$

with $T_{\min} = 0.7$ and $T_{\max} = 1.3$. Rejected samples are excluded from the pseudo-label loss. The ordinal loss $\mathcal{L}_{\mathrm{ord}}$ is not applied to pseudo-labeled samples; we therefore do not derive ordinal targets from $\tilde{p}(x)$ in our final configuration.

3.5 Hybrid Objective, Consistency, and Training Protocol

To align learning with clinical priorities, the supervised categorical term addresses class imbalance, the ordinal term encodes disease ordering, and the pseudo-label and consistency terms stabilize learning from confidently predicted, low-uncertainty samples. For labeled samples with ground-truth y, we use class-balanced focal loss $\mathcal{L}_{\text{cls}}$ with effective-number weights $\alpha_y = \frac{1-\beta_{\text{cb}}}{1-\beta_{\text{cb}}^{n_y}}$ (renormalized over classes), $\beta_{\text{cb}} = 0.99$, and $\gamma = 2$: For accepted pseudo-labeled samples j with $\tilde{p}_j$, we minimize cross-entropy to $\tilde{p}_j$ and add a prediction-consistency penalty between two independent stochastic augmentations $x_j^{(a)}, x_j^{(b)}$:

$$\begin{aligned}
\mathcal{L}_{\text{cls}} &= \frac{1}{|D_{\text{sup}}|} \sum_{i \in D_{\text{sup}}} -\tilde{\alpha}_{y_i} \, (1 - p_{i,y_i})^{\gamma} \log p_{i,y_i}, \\
\mathcal{L}_{\text{pl}} &= \frac{1}{|D_{\text{pl}}|} \sum_{j \in D_{\text{pl}}} -\sum_{k=1}^{K} \tilde{p}_{j,k} \log p_{j,k}, \\
\mathcal{L}_{\text{cons}} &= \frac{1}{2|D_{\text{pl}}|} \sum_{j \in D_{\text{pl}}} \Big(\mathrm{KL}\big(p(x_j^{(a)}) \,\|\, p(x_j^{(b)})\big) + \mathrm{KL}\big(p(x_j^{(b)}) \,\|\, p(x_j^{(a)})\big) \Big).
\end{aligned} \tag{13}$$

Both $\mathcal{L}_{\text{pl}}$ and $\mathcal{L}_{\text{cons}}$ operate on the categorical head p only. Augmentations are sampled independently per view from the same pipeline (resize, color jitter, rotation, blur, normalization; MixUp/CutMix disabled in Stage 2). During training, batch normalization layers remain trainable and compute per-batch statistics; no special sharing across the two views is used beyond standard mini-batch aggregation. The total objective is

$$\mathcal{L}_{\text{total}} = \lambda_{\text{cls}} \mathcal{L}_{\text{cls}} + \lambda_{\text{ord}} \mathcal{L}_{\text{ord}} + \lambda_{\text{cls}} \mathcal{L}_{\text{pl}} + \lambda_{\text{cons}} \mathcal{L}_{\text{cons}}, \tag{14}$$

with $\lambda_{\text{cls}} = 1.0$, $\lambda_{\text{ord}} = 0.5$, $\lambda_{\text{cons}} = 0.2$. Domains: $\mathcal{L}_{\text{cls}}$ on D_{sup}; $\mathcal{L}_{\text{ord}}$ on D_{sup} only; $\mathcal{L}_{\text{pl}}$ and $\mathcal{L}_{\text{cons}}$ on D_{pl}. The consistency weight ramps linearly from 0 to λ_{cons} over the first half of Stage 2.

4 Experiments

4.1 Experimental Settings

4.1.1 Dataset, Preprocessing, and Evaluation Protocol

We conduct internal validation on the APTOS 2019 Blindness Detection dataset for five-class diabetic retinopathy severity grading using single-modality RGB fundus photographs with image-level labels only. No external data, modalities, or auxiliary annotations are used. All results are reported on a single stratified 80/20 train/validation split ("canonical split") with a single random seed.

Evaluation protocol and metrics. Metrics are computed from the categorical head's softmax probabilities. The primary validation metric is quadratic weighted kappa (QWK), which captures ordinal agreement with labels (0–4). Secondary metrics are top-1 accuracy, macro-averaged F1, and macro one-vs-rest ROC-AUC (ovr). For AUC, if a class has zero positives in the validation split, its AUC is undefined and excluded from the macro average. Model selection (early stopping) uses QWK on the held-out validation set. We do not report multi-seed repeats, cross-validation, or statistical significance testing (e.g., bootstrap confidence intervals) in this submission.

Image preprocessing. Each image is center-cropped to remove black borders and peripheral artifacts and to standardize the field of view, then resized to 384×384 pixels before model input. The 384×384 resolution was chosen after a pilot sweep over 224/384/512 that indicated a favorable balance between sensitivity to small red lesions and computational cost. We do not report quantitative resolution ablations in this submission. Unless otherwise stated, we do not exclude images using automated quality filters; we log blur and illumination scores and defer sensitivity analysis to future work.

4.1.2 Aligned Methods and Implementation Details

We align all implementations and reported results with the final method described in Section 3. The design addresses sensitivity to microlesions, ordinal consistency, and robustness to noisy, imbalanced labels. The objective combines class-balanced focal loss for the categorical head with a focal-augmented CORAL loss for the ordinal head, using effective-number class weights. For pseudo-labeling, we apply uncertainty-aware gating based on mutual information (MI) from Monte Carlo dropout. A pseudo-label is accepted only if the maximum class probability exceeds a confidence threshold τ and the MI falls below a class-aware cutoff ν. Accepted pseudo-labels are sharpened with an uncertainty-dependent temperature, and we do not use entropy-only gating. For consistency, we add a lightweight prediction-consistency penalty across two stochastic augmentations for accepted pseudo-labeled samples and do not use an EMA teacher. For class imbalance, we adopt class-balanced focal loss without a WeightedRandomSampler; sampler-only and sampler-plus-class-balance configurations are not reported here.

Backbone and CSCA. The backbone is DenseNet-121 pretrained on ImageNet. We integrate the proposed cross-scale CSCA module, performing dual-pooling channel gating at each scale and deep-queried descriptor attention across levels, followed by 1×1 fusion. Global average pooling (GAP) over the fused feature produces the readout for dual heads: (i) a 5-way categorical softmax head for $p \in \Delta^5$, and (ii) a CORAL ordinal head outputting four cumulative logits for $q \in (0,1)^4$.

Optimization and schedule. We use AdamW ($\mathrm{lr} = 1\mathrm{e}{-4}$, weight decay $= 1\mathrm{e}{-4}$), mixed precision, and gradient clipping (max-norm 5.0). Stage 1 (supervised) trains for 50 epochs without pseudo-labels. We then run MI-based pseudo-labeling on the training set using T stochastic passes, accept and sharpen labels using (τ, ν), and proceed to Stage 2 (40 epochs) with consistency regularization. The consistency weight ramps linearly from 0 to its target over the first half of Stage 2. Augmentations include resize, color jitter, random rotation ($\pm15^\circ$), horizontal flip, Gaussian blur, and normalization to ImageNet mean/variance. MixUp (alpha 0.4) and CutMix (alpha 1.0) are used in Stage 1 only. Learning-rate schedules and all other training settings are shared across methods for fair comparison.

Fixed hyperparameters (APTOS). MC dropout passes $T = 10$; $\tau = 0.9$; class-wise MI cutoffs $\{\nu_c\}$ selected via per-class quantiles on the training fold; loss weights $\lambda_{\mathrm{cls}} = 1.0$, $\lambda_{\mathrm{ord}} = 0.5$, $\lambda_{\mathrm{cons}} = 0.2$; focal exponents $\gamma = 2$ (categorical) and $\gamma_{\mathrm{ord}} = 2$ (ordinal); effective-number parameter $\beta_{\mathrm{cb}} = 0.99$.

AUC computation details. All AUCs use the categorical head's probabilities in a one-vs-rest setup. Classes with zero positives in the validation split are excluded from the macro AUC. Per-class counts for the canonical split are provided in the released manifest; fold-wise exclusions do not apply since we do not report cross-validation here.

4.2 Main Performance comparisons

4.2.1 Baselines Setup

To evaluate the effectiveness of the proposed method, we compare it with several widely used convolutional neural network architectures for image classification. Specifically, we consider ResNet-50, VGG-16, Inception-V3, and MobileNet-V3 as baseline models. These networks represent different design paradigms, ranging from deep residual learning to lightweight mobile architectures.

All baseline models are trained under the same experimental protocol to ensure a fair comparison. Each network is optimized using identical training splits, preprocessing procedures, and evaluation metrics. We report multiple metrics commonly used for classification evaluation, including accuracy (ACC), precision (Prec), recall (Rec), F1 score, area under the ROC curve (AUC), and quadratic weighted kappa (QWK). The best-performing checkpoint for each model is selected according to validation performance.

4.2.2 Quantitative Results

We report validation performance on the canonical 80/20 stratified split for the proposed full model. Table 1 reports the quantitative comparison with several representative CNN architectures, including ResNet-50, VGG-16, Inception-V3, and MobileNet-V3. Among the baselines, VGG-16 achieves the

Table 1: Comparison with standard CNN architectures for image classification on APTOS 2019. All metrics except QWK are reported in percentage (%). The best results are highlighted in bold.

Method	ACC ↑	Prec ↑	Rec ↑	F1 ↑	AUC ↑	QWK ↑
ResNet-50	68.67	69.27	68.67	67.95	89.13	0.79
VGG-16	70.67	71.13	70.67	70.47	88.19	0.84
Inception-V3	68.00	69.49	68.00	67.18	88.71	0.80
MobileNet-V3	68.00	67.75	68.00	67.40	89.08	0.80
Ours	**73.33**	**74.84**	**73.33**	**73.23**	**91.54**	**0.85**

strongest performance with an accuracy of 70.67 % and a QWK score of 0.8408. Our method further improves the accuracy to 73.33 % and achieves the best QWK score of 0.8515. In addition, our model obtains the highest AUC (0.9154), surpassing the best baseline by more than 2 points. Overall, the proposed approach consistently outperforms all baseline architectures across most metrics, demonstrating the effectiveness of the proposed design for improving classification performance.

4.2.3 Visualization Comparison

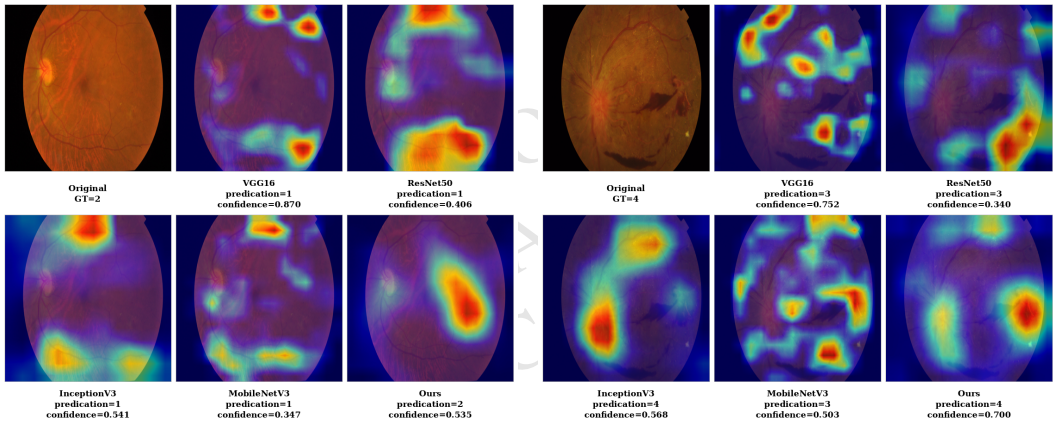

Figure 2: Visualization results from GradCAM between our method and the other baseline methods.

We further visualize class activation maps (CAMs) to examine the spatial attention of different models. Two representative fundus images are selected for comparison across several standard backbones and our method in Figure 2. In the first case, the image is labeled as grade 2. Most baseline models underestimate the severity and predict grade 1, with CAMs showing diffuse responses over large retinal regions. In contrast, our method correctly predicts grade 2 and produces more localized activations around lesion-relevant areas. In the second case, corresponding to grade 4, several baselines predict grade 3, while our model correctly identifies the highest severity level. The CAMs further indicate that our model focuses more consistently on pathological structures associated with severe disease. These visualizations suggest that the proposed model captures more disease-relevant cues, which helps mitigate the common tendency of underestimating DR severity.

4.3 Ablation Study

Table 2 presents the ablation study of the proposed components. Removing the CSCA module leads to a noticeable performance drop, reducing the accuracy from 73.33 % to 69.33 %, which confirms the importance of cross-scale feature selection for capturing subtle retinal lesions. When only partial channel attention mechanisms are used, the performance improves compared with the baseline but remains below the full model. For instance, the channel-gating-only variant achieves 72.67 % accuracy and a QWK score of 0.86, indicating that channel-wise attention contributes to discriminative feature learning but lacks the full cross-scale interaction modeled by CSCA. Finally, integrating CSCA with the uncertainty-gated pseudo-label training strategy yields the best results across most

Table 2: Ablation study of the proposed components. CSCA denotes the Cross-Scale Channel Attention module, and PL refers to the uncertainty-gated pseudo-labeling strategy. The best results are highlighted in bold.

Configuration	ACC ↑	Prec ↑	Rec ↑	F1 ↑	AUC ↑	QWK ↑
w/o CSCA, w/o PL	69.33	71.82	69.33	69.40	89.65	0.85
CSCA only	70.67	71.13	70.67	70.68	89.00	0.85
Channel gating only	72.67	**75.25**	72.67	72.71	90.50	**0.86**
Temporal descriptor only	72.00	73.95	72.00	71.96	89.44	0.84
CSCA + PL (Ours)	**73.33**	74.84	**73.33**	**73.23**	**91.54**	0.85

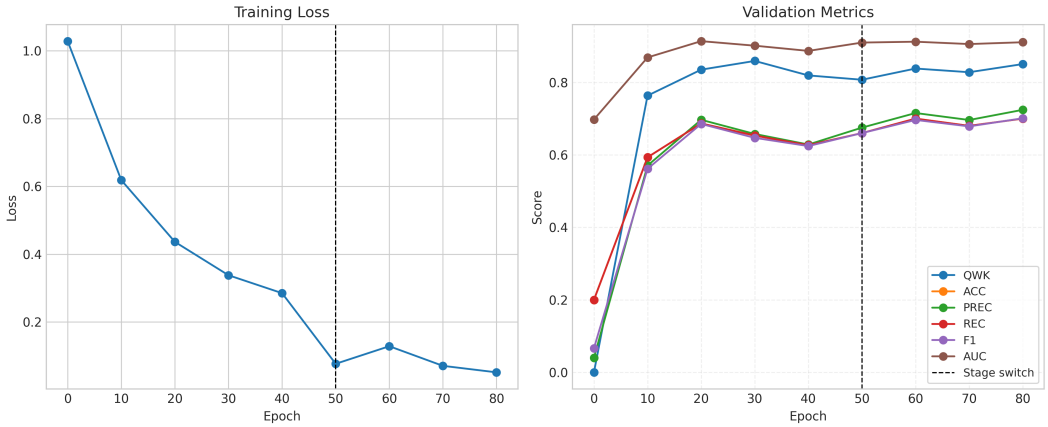

Figure 3: Learning curves across epochs: training loss (left axis) and validation accuracy (right axis) on the canonical 80/20 split. The dashed line marks the transition from Stage 1 (supervised) to Stage 2 (pseudo-labels with consistency).

metrics, achieving 73.33 % accuracy and the highest AUC of 91.54 %. These results demonstrate that both the cross-scale attention design and the pseudo-label training strategy contribute complementary benefits to the final model.

4.4 Complexity and Inference Efficiency

We measure complexity analytically and report the incremental overhead from CSCA relative to the DenseNet-121 backbone. For completeness, Table 3 lists exact base vs. base+CSCA counts at 384×384 (measured with fvcore; MACs rounded to two decimals). With $r = 4$, descriptor dimension $d = 256$, context width $C_u = 64$, fusion $C_f = 1024$, and $L = 4$ scales (channels $C_l \in \{256, 512, 1024, 1024\}$), CSCA introduces approximately 6.49 million parameters in total (channel gating ≈1.82M; descriptor projections ≈1.70M; context projection ≈0.016M; 1×1 fusion ≈2.95M). At 384×384 inputs, the dominant extra MACs arise from the 1×1 fusion at the deepest map resolution (about 0.425 G MACs), which is a small fraction of a DenseNet-121 forward at this resolution.

Model @384×384	Params (M)	MACs (G)
DenseNet-121 (base)	7.98	8.48
DenseNet-121 + CSCA (ours)	14.47	8.90

Table 3: DenseNet-121 base vs base+CSCA parameter and MAC counts at 384×384. CSCA adds ≈ 6.49M params and 0.43G MACs.

4.5 Learning Dynamics

Figure 3 illustrates the training loss and validation accuracy across epochs on the canonical split. During the first stage, the model exhibits stable optimization behavior, with a consistent decrease in training loss accompanied by gradual improvements in validation accuracy. This stage corresponds to supervised training using the hybrid objective, where the categorical and ordinal heads jointly guide feature learning. After pseudo-label refinement is introduced in Stage 2, the training dynamics remain stable while the validation accuracy continues to improve slightly. This suggests that the uncertainty-gated pseudo-labeling and the consistency regularization provide additional supervisory signals without destabilizing optimization. Overall, the two-stage training scheme leads to smooth convergence and consistent validation performance.

5 Conclusion

Automated grading of diabetic retinopathy from fundus photographs must detect sparse microlesions, respect the ordinal severity scale, and remain robust to imbalance and label noise. We present a single-stream, fundus-only model that integrates Cross-Scale Channel Attention with dual supervision from a categorical softmax head and a CORAL ordinal head. A hybrid focal-plus-ordinal loss encourages balanced, order-consistent learning, and uncertainty-gated pseudo-label refinement reduces noisy supervision. On the canonical stratified 80/20 APTOS 2019, the proposed system achieves a QWK of 0.85, an accuracy of 73.33%, a macro-F1 score of 73.23%, and a macro one-vs-rest AUC of 91.54%. These results demonstrate the effectiveness of the proposed cross-scale attention design and the uncertainty-gated pseudo-label training strategy for robust diabetic retinopathy grading. Future work will evaluate across independent cohorts and devices, calibrate probabilities via temperature scaling with ECE/Brier and referable-DR operating points, assess lesion-level saliency, and test additional and lightweight backbones.

6 Ethics Statement

This research study used only the publicly data made available 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

Julius Adebayo, Justin Gilmer, Ian Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pp. 9505–9515, 2018.

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

Qingyang Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks. In *Advances in Neural Information Processing Systems*, 2020.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Improved visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision*, pp. 839–847, 2018.

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*, pp. 9268–9277, 2019.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, 2016.

Early Treatment Diabetic Retinopathy Study Research Group. Grading diabetic retinopathy from stereoscopic color fundus photographs—an extension of the modified airlie house classification. *Ophthalmology*, 94(7):761–774, 1987. doi: 10.1016/S0161-6420(87)33593-8.

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.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

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*, pp. 770–778, 2016.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.

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*, pp. 4700–4708, 2017.

Peng-Tao Jiang, Qibin Zhang, Qibin Hou, Ming-Ming Cheng, Yunchao Wei, Hanfang Xiong, and Jianming Feng. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021. doi: 10.1109/TIP.2021.3079659.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pp. 6402–6413, 2017.

Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 510–519, 2019.

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*, pp. 2980–2988, 2017.

Zhen Niu, Mo Zhou, Liu Wang, and Xin Gao. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4920–4928, 2016.

Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 10–15, 2015.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

Daniel SW Ting, Carol Y 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.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020a.

Qilong Wang, Bottleneck Wu, Peng Hu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11531–11539, 2020b.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, pp. 3–19. Springer, 2018.

Qizhe Xie, Eduard Hovy, Nuno He, Minh-Thang Luong, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10687–10698, 2020.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929, 2016.