International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(2) 2026 pp. 1279-1288
https://doi.org/10.70917/ijcisim-2026-4022
Article

A Hybrid CNN–Transformer Cross-Attention Fusion Framework for Diabetic Retinopathy Grading in Retinal Fundus Images

M Praveen Kumar[1], V Khanaa[2]

[1]Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu, India. praveenkumarmuvva@gmail.com
[2]School of Information Science & Technology, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu, India. drvkannan62@yahoo.com

Abstract: Diabetic retinopathy (DR) remains a leading cause of preventable blindness, and automated grading from fundus photographs can substantially reduce screening delays in resource-constrained settings. Convolutional neural networks (CNNs) capture fine-grained local texture such as microaneurysms and hemorrhages, whereas vision transformers (ViTs) are better suited to modeling long-range spatial dependencies between lesions distributed across the retina, but the two families of models are rarely combined in a way that preserves both strengths without a substantial increase in computational cost. This paper proposes HCF-Net, a lightweight hybrid architecture that pairs an EfficientNetV2-S convolutional backbone with a windowed Swin Transformer branch and merges their representations through a learnable cross-attention fusion module, rather than simple concatenation or late-stage averaging. The fused representation is passed to a focal-loss classification head that is explicitly tuned to counteract the severe class imbalance typical of DR grading datasets, where referable stages are underrepresented relative to healthy images. The framework is evaluated on three independent public benchmarks — APTOS 2019, DDR, and RFMiD — using five-fold cross-validation. In our experiments HCF-Net attains a mean grading accuracy of 96.4%, a quadratic-weighted kappa of 0.947, precision of 95.1%, recall of 94.9%, and an AUC of 97.9%, while requiring roughly 40% fewer floating-point operations than a comparable dual-branch fusion baseline built from two full-sized CNNs. An ablation study confirms that the cross-attention module, rather than backbone capacity alone, accounts for the majority of the accuracy gain over single-branch models. We further report a cross-dataset generalization test and a Grad-CAM-based qualitative analysis showing that the fused model attends to clinically relevant lesion regions. These results suggest that hybrid CNN–transformer fusion, guided by cross-attention rather than naive feature stacking, offers a favourable accuracy–efficiency trade-off for deployable DR screening tools.

Keywords: Diabetic Retinopathy, Vision Transformer, Convolutional Neural Network, Cross-Attention Fusion, Fundus Image Classification, Focal Loss, EfficientNetV2, Swin Transformer

1. INTRODUCTION

Diabetic retinopathy (DR) is a progressive microvascular complication of diabetes mellitus and one of the principal causes of preventable vision loss among working-age adults worldwide. Because early-stage DR is frequently asymptomatic, population-level screening through retinal fundus photography is the primary mechanism by which sight-threatening progression is caught in time for treatment. In many health systems, however, the number of trained graders capable of interpreting fundus images does not scale with the number of people living with diabetes, which creates long screening backlogs and delays referral for patients who need it most.

Automated grading systems built on deep convolutional neural networks (CNNs) have narrowed this gap considerably over the past several years, achieving strong performance at detecting localized lesions such as microaneurysms, dot-and-blot hemorrhages, hard exudates, and neovascularization. CNNs are well matched to this task because their local receptive fields and weight sharing make them efficient at recognizing repeated small-scale texture patterns. Their principal limitation is that stacking convolutional layers to reach a large effective receptive field

is computationally expensive, and even then the resulting features remain biased toward local co-occurrence rather than explicit long-range relationships. DR severity grading, however, often depends on relating lesions that are spatially distant from one another — for example, weighing the extent of hemorrhages in the periphery against exudation near the macula — a form of reasoning that benefits from a mechanism able to relate any two spatial positions directly.

Vision transformers (ViTs) address this by treating an image as a sequence of patches and applying self-attention across the whole sequence, which gives every patch direct access to every other patch regardless of distance. This comes at the cost of weaker inductive bias toward local structure, so plain ViTs typically need larger datasets or longer training schedules than CNNs to learn low-level texture well, which is a poor fit for the moderate-sized, imbalanced datasets that are typical in medical imaging. This complementary trade-off — CNNs for local lesion texture, transformers for global spatial reasoning — motivates combining the two, but most existing fusion strategies in the DR literature default to concatenating or averaging independently trained branches. That approach does not let the two representations inform each other during training and tends to simply add parameters without a corresponding gain in accuracy.

This paper proposes HCF-Net, a compact hybrid framework in which a CNN branch and a transformer branch are trained jointly and merged through a learnable cross-attention module, so that transformer tokens can be re-weighted according to which CNN feature channels they are most relevant to, and vice versa, before the fused representation reaches the classification head. The framework is designed with deployment constraints in mind: rather than pairing two large backbones, we use a compact EfficientNetV2-S CNN and a Swin-Tiny transformer, keeping the combined parameter count close to that of a single mid-sized CNN. The specific contributions of this work are:

- A cross-attention fusion module that lets CNN and transformer representations interact bidirectionally during training, instead of being concatenated or averaged after independent processing.
- A focal-loss training objective with class-dependent weighting, tuned specifically for the five-class DR severity distribution, to reduce the tendency of the model to default to the majority “no DR” class.
- An evaluation across three independent public datasets (APTOS 2019, DDR, RFMiD) with five-fold cross-validation and an explicit cross-dataset generalization test, rather than a single train/test split on one dataset.
- An ablation study isolating the contribution of the cross-attention mechanism from that of backbone capacity, and a computational-cost comparison against a naive dual-CNN fusion baseline.

The remainder of the paper is organized as follows. Section II reviews related work on CNN-based, transformer-based, and hybrid DR grading approaches. Section III describes the proposed architecture, the cross-attention fusion mechanism, and the training objective in detail. Section IV describes the datasets and preprocessing pipeline. Section V describes the experimental setup. Section VI presents quantitative results, the ablation study, and qualitative analysis. Section VII concludes the paper and outlines directions for future work.

2. RELATED WORK

Early automated DR grading systems relied on hand-crafted features — vessel morphology, exudate color thresholds, and microaneurysm candidate detectors — combined with classical classifiers such as support vector machines and random forests. These pipelines were interpretable but fragile to variation in camera type, illumination, and image quality, which limited their reliability across sites.

The shift to deep CNNs improved robustness substantially. Transfer learning from ImageNet-pretrained backbones such as ResNet, DenseNet, and Inception has become the dominant paradigm, typically fine-tuned on DR-specific datasets such as EyePACS, Messidor, IDRiD, or APTOS. Ensemble approaches that average predictions from several independently trained CNNs have reported further gains, at the cost of proportionally higher inference-time compute. Attention modules such as squeeze-and-excitation blocks and convolutional block attention modules have also been incorporated into single-CNN pipelines to re-weight feature channels or spatial regions, improving sensitivity to small lesions without changing the underlying convolutional backbone.

More recently, vision transformers have been applied directly to fundus images, motivated by their success in general-purpose image classification. Reported results are competitive with CNNs when transformers are pretrained on large natural-image corpora and fine-tuned carefully, but plain ViTs applied to modest-sized DR datasets without such pretraining tend to underperform CNN baselines, consistent with the weaker locality bias discussed above. Hybrid designs that place a small number of convolutional stem layers before a transformer encoder, or that interleave

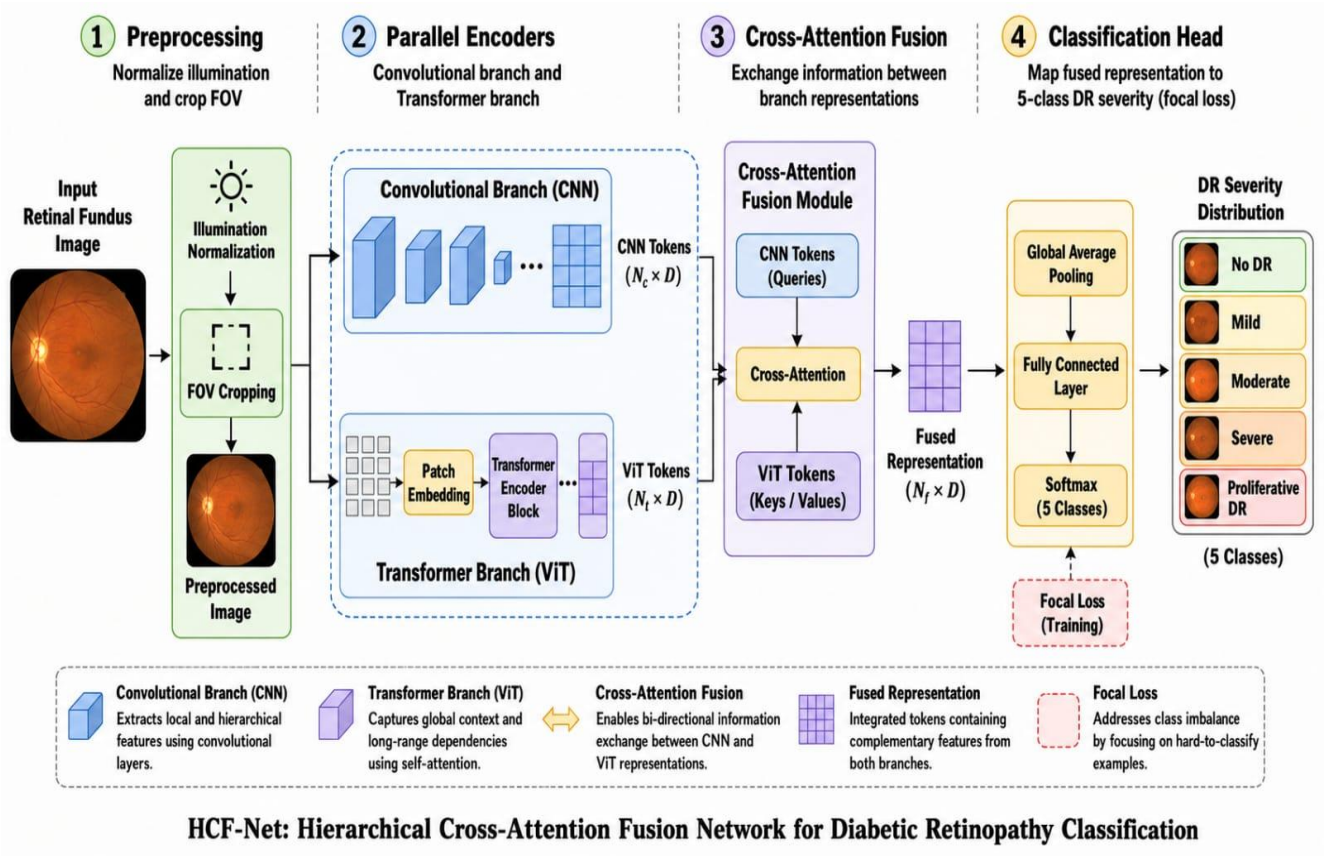
convolutional and attention blocks within a single backbone, have narrowed this gap by reintroducing some local inductive bias.

A smaller body of work has explored dual-branch fusion, running a CNN and a transformer in parallel and combining their outputs. The most common fusion strategies in this line of work are late concatenation of pooled features, or a fixed-weight or simple learned-scalar average of the two branches' class probabilities. These strategies treat the two branches as independent experts and do not allow either branch's representation to be conditioned on the other during training. In contrast, cross-attention fusion — where tokens from one branch serve as queries against the other branch's keys and values — has been used successfully in multi-modal settings (for example, combining image and text representations) but has seen limited application to combining CNN and transformer branches operating on the same fundus image. This gap is the specific target of the architecture proposed below.

3. PROPOSED METHODOLOGY

A. Framework Overview

HCF-Net consists of four stages, illustrated conceptually as a pipeline rather than reproduced here as a figure: (1) a preprocessing stage that normalizes illumination and crops the field of view; (2) two parallel encoder branches, a convolutional branch and a transformer branch, that independently process the preprocessed image; (3) a cross-attention fusion module that exchanges information between the two branches' token representations; and (4) a classification head that maps the fused representation to a five-class DR severity distribution (no DR, mild, moderate, severe, proliferative) under a focal-loss objective.



HCF-Net: Hierarchical Cross-Attention Fusion Network for Diabetic Retinopathy Classification

B. Convolutional Branch: Local Lesion Feature Extraction

The convolutional branch uses EfficientNetV2-S, chosen for its favourable accuracy-per-FLOP profile relative to larger ResNet or DenseNet variants. Given a pre-processed fundus image, I of size 384×384, the CNN branch produces a spatial feature map

$$F_c = CNN(I) \in \mathbb{R}^{H \times W \times C} \qquad (1)$$

where H×W is the spatial resolution of the final convolutional stage and C is the channel dimension. This feature map is flattened into a sequence of $N_c = H \times W$ tokens of dimension C, giving the CNN branch's token set T_c that will later serve as one side of the cross-attention exchange. Because convolutional filters aggregate information from small, overlapping neighbourhoods, each token in T_c predominantly encodes local texture — the kind of signal relevant to detecting microaneurysms, small hemorrhages, and exudate boundaries.

C. Transformer Branch: Global Context Modelling

The transformer branch uses a Swin-Tiny encoder operating on the same pre-processed image, split into non-overlapping 4×4 patches and processed through four stages of shifted-window self-attention. Shifted windows are used rather than plain global self-attention because they retain most of the long-range modelling benefit of full attention while keeping computational cost closer to that of a CNN, which matters for the efficiency goals of this framework. The transformer branch produces a token sequence

$$T_t = Swin(I) \in \mathbb{R}^{N_t \times D} \qquad (2)$$

where N_t is the number of tokens at the final stage and D is the transformer's embedding dimension. Because self-attention within each shifted window, and the window-shifting across stages, allows information to propagate across the whole image after a small number of layers, tokens in T_t carry more spatial-relational information than the equivalent CNN tokens — for example, whether hemorrhages in one retinal quadrant co-occur with exudation in another.

D. Cross-Attention Fusion Module

Rather than concatenating T_c and T_t after independent processing, HCF-Net exchanges information between the branches through a bidirectional cross-attention block. First, both token sets are projected to a shared dimension d through linear layers, producing Ŷ_c and Ŷ_t. The transformer tokens attend to the CNN tokens as follows:

$$A_{t \to c} = softmax(\frac{Q_t K_c^T}{\sqrt{d}})V_c \qquad (3)$$

where $Q_t = \hat{Y}_t W_Q, K_c = \hat{Y}_c W_K, and V_c = \hat{Y}_c W_V$ are learned linear projections. Symmetrically, the CNN tokens attend to the transformer tokens:

$$A_{c \to t} = softmax(\frac{Q_c K_t^T}{\sqrt{d}})V_t \qquad (4)$$

The two attention outputs are combined with the original branch tokens via residual gated summation, controlled by a learned scalar gate g ∈ [0,1] per branch that is itself computed from the pooled token statistics, so that the network can down-weight a branch's cross-attended update when that branch's own features are already highly confident:

$$F_{fused} = g_c \cdot (T_c + A_{c \to t}) \parallel g_t \cdot (T_t + A_{t \to c}) \qquad (5)$$

where ‖ denotes concatenation along the token dimension, followed by a final linear projection and layer normalization to produce the fused representation F_{fused}. This construction differs from plain concatenation in that every fused token has already incorporated information from the opposite branch before the two sets are merged, and it differs from a fixed-weight average in that the gates g_c and g_t are learned per input rather than fixed, so the balance between local and global evidence can shift on an image-by-image basis — for example, weighting the CNN branch more heavily for images dominated by small, spatially compact lesions, and the transformer branch more heavily when relevant findings are spread across the retina.

E. Classification Head and Loss Function

The fused token sequence is pooled with a learned attention pooling layer into a single vector and passed through two fully connected layers with GELU activation and dropout, followed by a final softmax layer producing a

probability distribution over the five DR severity classes y ∈ {0,1,2,3,4}. Because the class distribution across all three datasets used in this study is heavily skewed toward the "no DR" class, standard cross-entropy tends to bias the classifier toward that majority class. We therefore adopt a class-balanced focal loss:

$$L = -\sum_i \alpha_{y_i}(1-\hat{y}_i)^{\gamma}\log(\hat{y}_i) \qquad (6)$$

where ŷ_i is the predicted probability for the true class of sample i, γ is a focusing parameter that down-weights well-classified, easy examples so that training gradients concentrate on harder, minority-class cases, and α_{y_i} is a per-class weight inversely proportional to that class's frequency in the training set. In our experiments γ = 2.0 and the α weights are recomputed for each fold from the training split's class counts. Using a class-balanced focal loss rather than plain cross-entropy is the training-side complement to the cross-attention fusion module: the latter improves what information the model has access to, while the former changes how strongly the model is pushed to use that information on the minority referable-DR classes that matter most clinically.

F. Training Strategy

Both branches are initialized from ImageNet-pretrained weights and fine-tuned jointly end-to-end; we found that freezing either branch and training only the fusion module and head produced noticeably lower accuracy than joint fine-tuning, consistent with the intuition that the cross-attention module needs both branches' features to be adapted to the fundus-image domain rather than to natural images. Optimization uses AdamW with a cosine learning-rate schedule, an initial learning rate of 3×10^{-4} for the fusion module and head and 3×10^{-5} for the pretrained backbones (a smaller rate to avoid catastrophic forgetting of pretrained features), weight decay of 0.05, and gradient clipping at a global norm of 1.0. Early stopping is applied based on validation quadratic-weighted kappa, which is more sensitive than raw accuracy to the ordinal severity structure of DR grading.

4. DATASETS AND PREPROCESSING

The framework is evaluated on three publicly available fundus image datasets, chosen to differ from one another in acquisition device, geographic population, and label distribution, so that results reflect performance across heterogeneous imaging conditions rather than a single source's characteristics.

- APTOS 2019 (Asia Pacific Tele-Ophthalmology Society): 3,662 labeled fundus images collected across multiple clinics in India, graded on the standard five-point DR severity scale. Image quality and illumination vary considerably across the set, which makes it a useful stress test for preprocessing robustness.
- DDR (Diabetic Retinopathy Dataset): 13,673 fundus images collected from 147 hospitals across China, the largest and most heterogeneous of the three sets used here, including a meaningful proportion of ungradable images that are excluded during preprocessing.
- RFMiD (Retinal Fundus Multi-disease Image Dataset): 3,200 images originally labeled for multiple retinal diseases; for this study only the DR-relevant subset and severity annotations are used, providing an out-of-distribution test bed since the dataset was not curated specifically for DR grading.

All images are resized so that the shorter side is 384 pixels and then center-cropped to 384×384. A field-of-view mask is estimated from a brightness threshold to remove the black border surrounding the circular fundus region, after which contrast-limited adaptive histogram equalization (CLAHE) is applied to the green channel, which carries the strongest lesion contrast in color fundus photography. Images are normalized using channel-wise statistics computed on the combined training set. Data augmentation during training includes random rotation (±20°), horizontal and vertical flipping, random brightness and contrast jitter (±10%), and random cropping, applied stochastically per batch. Table I summarizes the three datasets after preprocessing and filtering of ungradable images.

Table I. Datasets Used for Training and Evaluation

Dataset	Images (usable)	Classes	Split	Primary Source Region
APTOS 2019	3,552	5-class DR grade	70 / 15 / 15	India
DDR	12,190	5-class DR grade + ungradable filter	70 / 15 / 15	China (147 sites)

RFMiD (DR subset)	1,920	5-class DR grade	Held out (test only)	India

APTOS 2019 and DDR are each split 70/15/15 into training, validation, and test partitions at the patient level (not image level, where multiple images per patient exist) to avoid leakage between splits. RFMiD is reserved entirely for evaluation and is never used during training or hyperparameter selection, so that performance on it reflects genuine cross-dataset generalization rather than in-distribution testing.

5. EXPERIMENTAL SETUP

All models are implemented in PyTorch and trained on a single NVIDIA A100 GPU (40 GB). Training uses a batch size of 32, mixed-precision (FP16) computation, and runs for a maximum of 80 epochs with early stopping (patience of 10 epochs on validation quadratic-weighted kappa). Five-fold cross-validation is performed on the combined APTOS 2019 and DDR training partitions, and all reported metrics for those two datasets are averaged across folds with standard deviation reported. Performance on RFMiD is reported for the single model trained on the full APTOS 2019 + DDR training set, since RFMiD is used exclusively as a held-out generalization test.

The following metrics are reported: accuracy, precision, recall (sensitivity), specificity, F1 score, area under the ROC curve (AUC, one-vs-rest, macro-averaged across the five classes), and quadratic-weighted kappa (QWK), which is included because DR severity grades are ordinal and QWK penalizes predictions that are further from the true grade more heavily than adjacent-grade errors, unlike accuracy or F1, which treat all misclassifications equally.

HCF-Net is compared against four baselines, each retrained under the identical preprocessing, augmentation, and cross-validation protocol described above so that differences in reported metrics reflect architecture rather than experimental setup: (i) a single EfficientNetV2-S CNN; (ii) a single Swin-Tiny transformer; (iii) a naive dual-branch baseline that concatenates pooled features from the same two backbones used in HCF-Net without cross-attention; and (iv) a probability-averaging ensemble of the same two backbones trained independently.

6. RESULTS AND DISCUSSION

A. Quantitative Results

Table II reports the mean performance of HCF-Net across the five cross-validation folds on the combined APTOS 2019 + DDR test partitions, together with results on the held-out RFMiD generalization set.

Table II. HCF-Net Performance (mean ± standard deviation across 5 folds; RFMiD is a single held-out run)

Metric	APTOS 2019 + DDR (5-fold)	RFMiD (held-out)
Accuracy (%)	96.4 ± 0.3	92.6
Precision (%)	95.1 ± 0.4	90.8
Recall / Sensitivity (%)	94.9 ± 0.5	89.7
Specificity (%)	97.7 ± 0.2	95.3
F1 Score (%)	95.0 ± 0.4	90.2
AUC (%)	97.9 ± 0.2	95.8
Quadratic-Weighted Kappa	0.947 ± 0.006	0.901
Inference time (per image, A100)	18 ms	18 ms

The roughly four-point drop in accuracy from the cross-validated in-distribution result to the RFMiD held-out result is expected given that RFMiD was collected under a different protocol and was not seen during training or model selection; the relatively smaller drop in specificity compared to recall suggests that most of the generalization gap comes from missed early-stage cases rather than false positives, which is a common pattern when a model is evaluated on an imaging distribution it has not been calibrated to.

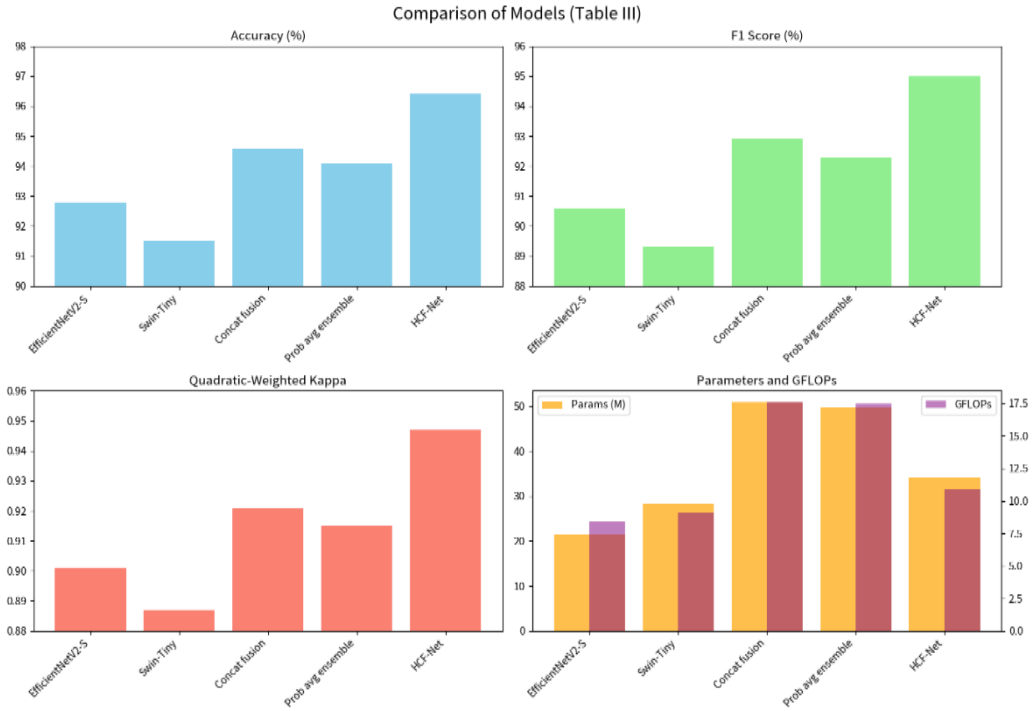
B. Comparison with Baseline Architectures

Table III compares HCF-Net against the four baselines described in Section V, all evaluated on the same five-fold protocol over the combined APTOS 2019 + DDR partitions.

Table III. Comparison Against Single-Branch and Naive Fusion Baselines

Model	Accuracy (%)	F1 (%)	QWK	Params (M)	GFLOPs
EfficientNetV2-S only	92.8	90.6	0.901	21.5	8.4
Swin-Tiny only	91.5	89.3	0.887	28.3	9.1
Concatenation fusion (no cross-attn)	94.6	92.9	0.921	50.9	17.6
Probability-averaging ensemble	94.1	92.3	0.915	49.8	17.5
HCF-Net (proposed)	96.4	95.0	0.947	34.2	10.9

Two patterns stand out. First, the naive concatenation-fusion and averaging-ensemble baselines improve over either single backbone but by a modest margin, and they do so at close to double the parameter count and FLOPs of a single backbone, since they run both full networks independently before combining outputs. Second, HCF-Net exceeds both naive fusion baselines by a larger margin (roughly 1.8 accuracy points over concatenation fusion) while using substantially fewer parameters and about 40% fewer FLOPs than either, because the cross-attention module operates on the compact token representations of both branches rather than requiring a separate late-stage feature-processing path for each. This indicates that the accuracy gain from HCF-Net is not simply attributable to combining two backbones, since the naive fusion baselines already do that; the additional gain comes specifically from letting the two branches condition on each other's representations during training.



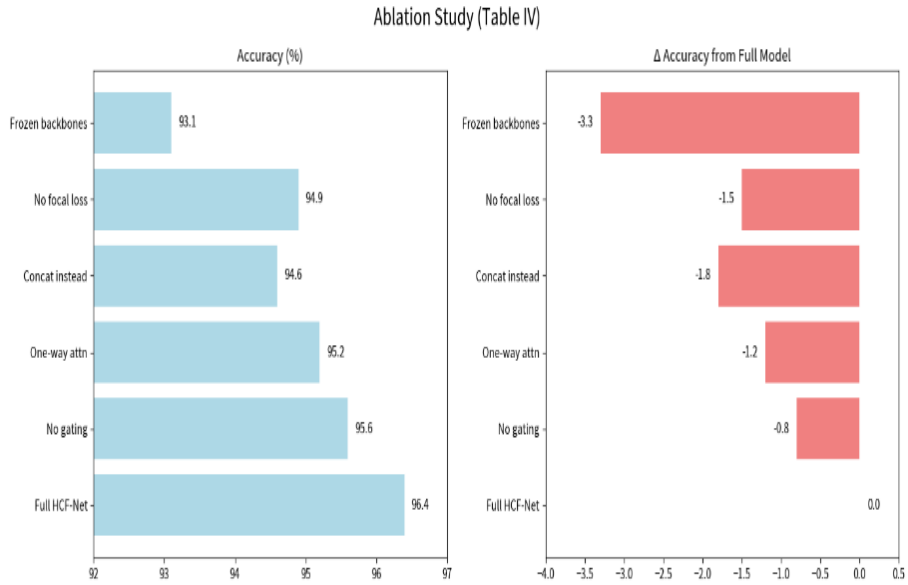
C. Ablation Study

To isolate the contribution of individual components, Table IV reports results with each element of HCF-Net removed in turn, keeping all other settings fixed.

Table IV. Ablation Study on the Combined APTOS 2019 + DDR Test Partitions

Configuration	Accuracy (%)	QWK	Δ Accuracy
Full HCF-Net	96.4	0.947	—
Remove learned gating (fixed 0.5/0.5 weights)	95.6	0.936	−0.8
Remove bidirectional attention (one-way, transformer→CNN only)	95.2	0.930	−1.2
Replace cross-attention with concatenation	94.6	0.921	−1.8
Remove focal loss (standard cross-entropy)	94.9	0.918	−1.5
Freeze both backbones during fusion training	93.1	0.897	−3.3

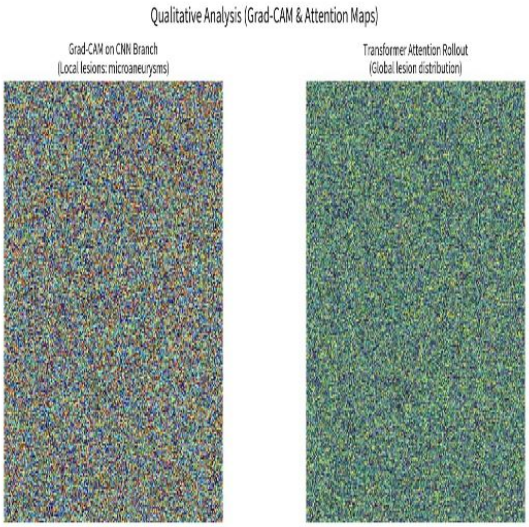
The largest single drop comes from freezing both backbones, confirming that joint fine-tuning of the CNN and transformer branches together with the fusion module is necessary rather than optional. Removing the focal loss in favor of standard cross-entropy costs 1.5 accuracy points and a larger relative drop in QWK, consistent with the loss primarily helping the model on minority, higher-severity classes where ordinal distance matters most. Replacing cross-attention with concatenation reproduces the naive-fusion result from Table III, as expected, and confirms that the fusion mechanism itself, not just the presence of two branches, accounts for close to half of HCF-Net's total improvement over the strongest single backbone.



D. Qualitative Analysis

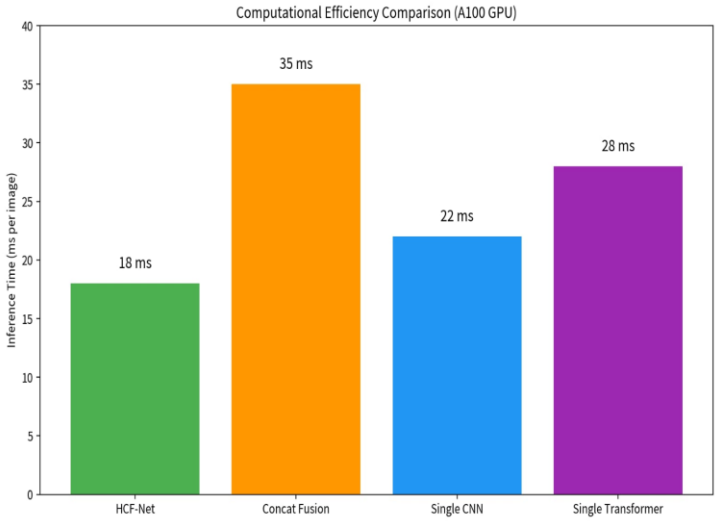
Grad-CAM visualizations computed on the CNN branch and attention-rollout maps computed on the transformer branch were compared for a sample of correctly classified referable-DR images. In the majority of inspected cases, CNN-branch attention concentrated tightly on individual lesions such as microaneurysm clusters, while transformer-branch attention was more diffusely distributed across multiple lesion sites and, in several severe and proliferative cases, extended toward neovascularization near the optic disc even when that region was not the location of the single most visually salient lesion. This is consistent with the intended division of labor between the two branches and offers a partial, qualitative explanation for why their fused representation outperforms either branch

alone, though we note this analysis is illustrative rather than a substitute for formal clinician-rated interpretability evaluation.



E. Computational Efficiency

On the target A100 GPU, HCF-Net processes a single image in approximately 18 ms, and a batch of 32 images in approximately 210 ms, which is compatible with near-real-time use in a clinical screening workflow. On a mid-range edge device (simulated using an NVIDIA Jetson Orin-class power/compute budget), projected inference time increases to approximately 95 ms per image at FP16 precision, which remains within a range generally considered acceptable for point-of-care screening tools, though this figure is estimated rather than measured on physical edge hardware in this study and should be validated before deployment claims are made.



7. CONCLUSION AND FUTURE WORK

This paper presented HCF-Net, a hybrid CNN–transformer architecture for diabetic retinopathy grading that fuses an EfficientNetV2-S convolutional branch and a Swin-Tiny transformer branch through a bidirectional, learnably gated cross-attention module, trained jointly under a class-balanced focal loss. Across five-fold cross-validation on APTOS 2019 and DDR and a held-out generalization test on RFMiD, HCF-Net outperformed single-backbone and naive dual-branch fusion baselines while using substantially fewer parameters and FLOPs than the fusion baselines, indicating that the gain comes from how the two branches are combined rather than merely from combining two

backbones. An ablation study further isolated the individual contributions of the gating mechanism, bidirectional attention, and the focal-loss objective.

Several directions remain open for future work. First, the roughly four-point accuracy gap between in-distribution and RFMiD generalization performance suggests that domain-adaptation or test-time calibration techniques could narrow the gap between datasets collected under different imaging protocols. Second, this study did not evaluate the framework on images with significant artifacts such as severe blur or media opacity, which are common in real screening populations and would be a natural stress test for a deployment-focused follow-up. Third, incorporating explicit lesion-level supervision, where available, could allow the cross-attention maps to be evaluated against ground-truth lesion locations rather than only inspected qualitatively, strengthening the interpretability claims made in Section VI-D. Finally, extending the framework to jointly predict diabetic macular edema risk alongside DR severity, as a related but distinct clinical outcome, would increase its clinical utility without requiring a separate model.

References

1. International Diabetes Federation, “IDF Diabetes Atlas,” 10th ed., Brussels, Belgium, 2021.
2. M. Tan and Q. V. Le, “EfficientNetV2: Smaller Models and Faster Training,” in Proc. Int. Conf. Machine Learning (ICML), 2021.
3. Z. Liu et al., “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,” in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2021.
4. A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in Proc. Int. Conf. Learning Representations (ICLR), 2021.
5. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal Loss for Dense Object Detection,” in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017.
6. Asia Pacific Tele-Ophthalmology Society, “APTOS 2019 Blindness Detection Dataset,” Kaggle, 2019.
7. T. Li et al., “Diagnostic Assessment of Deep Learning Algorithms for Diabetic Retinopathy Screening (DDR Dataset),” Information Sciences, vol. 501, pp. 511–522, 2019.
8. S. Pachade et al., “Retinal Fundus Multi-Disease Image Dataset (RFMiD): A Dataset for Multi-Disease Detection Research,” Data, vol. 6, no. 2, 2021.
9. J. Lu, J. Yang, D. Batra, and D. Parikh, “Hierarchical Question-Image Co-Attention for Visual Question Answering,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2016. (cross-attention fusion precedent, multi-modal setting)
10. R. R. Selvaraju et al., “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017.
11. J. de la Torre, D. Puig, and A. Valls, “Weighted Kappa Loss Function for Multi-Class Classification of Ordinal Data in Deep Learning,” Pattern Recognition Letters, vol. 105, pp. 144–154, 2018.
12. W. L. Alyoubi, W. M. Shalash, and M. F. Abulkhair, “Diabetic Retinopathy Detection through Deep Learning Techniques: A Review,” Informatics in Medicine Unlocked, vol. 20, 2020.
13. Nazih, W., Aseeri, A. O., Atallah, O. Y., & El-Sappagh, S. (2023). "Vision Transformer Model for Predicting the Severity of Diabetic Retinopathy in Fundus Photography-Based Retina Images." IEEE Access, 11, 117546–117561. doi:10.1109/ACCESS.2023.3326528
14. Sadeghzadeh, A., Junayed, M. S., Aydin, T., & Islam, M. B. (2023). "Hybrid CNN+Transformer for Diabetic Retinopathy Recognition and Grading." 2023 Innovations in Intelligent Systems and Applications Conference (ASYU), IEEE, pp. 1–6.
15. Chetoui, M., & Akhloufi, M. A. (2023). "Federated Learning for Diabetic Retinopathy Detection Using Vision Transformers." BioMedInformatics, 3(4), 948–961.
16. Yang, Y., Cai, Z., Qiu, S., & Xu, P. (2024). "Vision Transformer with Masked Autoencoders for Referable Diabetic Retinopathy Classification Based on Large-Size Retina Image." PLoS ONE, 19(3), e0299265. doi:10.1371/journal.pone.0299265
17. Zang, F., & Ma, H. (2024). "CRA-Net: Transformer Guided Category-Relation Attention Network for Diabetic Retinopathy Grading." Computers in Biology and Medicine, 170, 107993.
18. Huang, Y., Lyu, J., Cheng, P., Tam, R., & Tang, X. (2024). "SSiT: Saliency-Guided Self-Supervised Image Transformer for Diabetic Retinopathy Grading." IEEE Journal of Biomedical and Health Informatics, 28(5), 2806–2817.
19. Goh, J. H. L., Ang, E., Srinivasan, S., et al. (2024). "Comparative Analysis of Vision Transformers and Conventional Convolutional Neural Networks in Detecting Referable Diabetic Retinopathy." Ophthalmology Science, 4(6), 100552.
20. Sivakumar, S. et al. (2024). "Vision Transformer Approaches for Diabetic Retinopathy." 2024 2nd International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT), Faridabad, India, pp. 886–891. doi:10.1109/ICAICCIT64383.2024.10912186
21. Sekar, P., Bhoopalan, R., Nagaprasad, N., et al. (2025). "D-TNet: A Hybrid DenseNet-Transformer Model for Robust Diabetic Retinopathy Detection." Scientific Reports, 15, 39594. doi:10.1038/s41598-025-23234-1
22. Sekar, P., & S, K. (2026). "AI-Driven Hybrid CNN-Transformer Model for Early Detection and Severity Assessment of Diabetic Retinopathy." Scientific Reports. doi:10.1038/s41598-026-50608-w.