Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.076800

ARTICLE

Position-Wise Attention-Enhanced Vision Transformer for Diabetic Retinopathy Grading

Yan-Hao Huang* **and Yu-Tse Huang**

Department of Green Energy and Information Technology, National Taitung University, Taitung, Taiwan

*Corresponding Author: Yan-Hao Huang. Email: yhhuang@nttu.edu.tw

Received: 26 November 2025; Accepted: 19 January 2026; Published: 09 April 2026

ABSTRACT: Diabetic Retinopathy (DR) is a common microvascular complication of diabetes that progressively damages the retinal blood vessels and, without timely treatment, can lead to irreversible vision loss. In clinical practice, DR is typically diagnosed by ophthalmologists through visual inspection of fundus images, a process that is time-consuming and prone to inter- and intra-observer variability. Recent advances in artificial intelligence, particularly Convolutional Neural Networks (CNNs) and Transformer-based models, have shown strong potential for automated medical image classification and decision support. In this study, we propose a Position-Wise Attention-Enhanced Vision Transformer (PWAE-ViT), which integrates a positional attention module into the standard ViT architecture to strengthen spatial positional information and feature representation across image patches. The proposed module encourages the network to better model local and global contextual relationships, thereby improving DR grading performance. To evaluate the robustness of our model, experiments were conducted on two public retinal fundus image datasets: APTOS-2019 and Indian Diabetic Retinopathy Image Dataset (IDRiD). The proposed PWAE-ViT consistently outperforms the baseline ViT model, achieving classification accuracies of 84% and 62% on the APTOS-2019 and IDRiD datasets, respectively. These results demonstrate more accurate and reliable DR severity classification, offering a promising tool to assist clinicians in screening and diagnosis.

KEYWORDS: Transformer; attention mechanisms; medical imaging; diabetic retinopathy

1 Introduction

Diabetic retinopathy (DR) is a microvascular complication of diabetes. Approximately one-third of diabetic patients will develop some degree of retinopathy during their lifetime [1]. Because retinal damage is irreversible, advanced DR can ultimately lead to severe, permanent vision loss and is now one of the leading causes of blindness worldwide. Early DR is usually asymptomatic and therefore easily overlooked. By the time patients experience blurred vision and seek medical care, the disease is often already in a more advanced stage. Consequently, regular retinal screening is crucial for timely detection and treatment in diabetic populations.

With global population aging and lifestyle changes, the prevalence of diabetes is expected to continue rising. Currently, DR is primarily diagnosed by ophthalmologists through manual inspection of retinal fundus images. In the future, a shortage of specialists and growing screening demands may increase the risk of subjective judgment errors and misdiagnosis [2]. Establishing an objective, accurate, and scalable DR grading system is thus of great clinical value: it can alleviate the burden on clinicians and provide consistent, real-time decision support to accelerate referral and treatment of suspected cases.

According to international standards [3], DR severity is typically categorized into five stages: no DR (healthy retina), mild non-proliferative DR (mild NPDR), moderate NPDR, severe NPDR, and proliferative DR (PDR). These grades are determined by observing characteristic lesions in retinal fundus images [4]. Early-stage DR is primarily identified by microaneurysms. As the disease progresses, hemorrhages, soft exudates, and hard exudates appear, which manifest as red or white dots and localized white patches. In the most severe stage, PDR, abnormal neovascularization is observed. Given these visual manifestations, an automated DR grading system is a practical tool for assisting diagnosis. Although it cannot replace clinicians, such a system can provide objective evidence, reduce misdiagnosis, and substantially shorten diagnosis time.

In recent years, rapid progress in computer vision has driven its adoption in many aspects of daily life. Convolutional Neural Networks (CNNs) are among the most widely used architectures, supporting tasks such as semantic segmentation, image classification, and object detection. In medical imaging, these techniques have been applied to classify disease severity and to segment critical anatomical structures, such as blood vessels and organs, to reveal more detailed pathological information. In [5], the authors first enhanced images to emphasize vascular structures, then employed two pre-trained CNN models, VGG19 and InceptionV3, for transfer learning and compared their DR grading performance. In [6], a symmetric CNN architecture was proposed, incorporating two pooling layers at the network front as feature filters and performing DR grade classification in the fully connected layers.

Despite their foundational success, CNNs are constrained by their reliance on local receptive fields, which often limits their ability to capture global contextual information. In the context of DR, where lesions can be widely dispersed and the fundus background remains relatively uniform, CNNs may overlook subtle pathological regions, thereby increasing diagnostic error rates. In response to these limitations, the Transformer architecture [7] has redefined computer vision through the Vision Transformers (ViTs). By decomposing images into fixed-size patches and employing multi-head self-attention, ViT models long-range spatial dependencies that convolutional kernels often fail to capture. Recent years have witnessed a paradigm shift in retinal analysis, with Transformers demonstrating exceptional capacity for capturing long-range spatial dependencies compared to traditional CNNs [8]. Extensive benchmarks have shown that ViTs can significantly outperform conventional CNN architectures in detecting referable DR, particularly when dealing with high-resolution fundus images and complex pathological patterns [9].

Despite its potential, applying ViT to DR grading is hindered by the high data requirements noted in the original study [10] and the severe class imbalance inherent in clinical datasets. These factors often lead to the misclassification of rare but critical stages. Furthermore, while current ViT-based DR models have shown promise, they predominantly compute self-attention within encoder stages, overlooking fine-grained enhancement at the initial input. To bridge this gap, we propose PWAE-ViT, which introduces a parallel Position-dependent Weighting Attention Block (PWAB) that operates in tandem with the patch embedding layer. By fusing input-level attention cues with discretized tokens, our approach enriches feature representations prior to global processing, significantly enhancing sensitivity to subtle diabetic lesions.

Motivated by the above observations, this study focuses on medical image-based DR grading using computer vision techniques. To address the dual challenges of subtle lesion recognition and class imbalance in automated DR grading, we propose a method built on the ViTs backbone. Specifically, we introduce a position-aware self-attention module in parallel with the original front-end patch embedding stage to enhance local spatial cues and fine-grained feature representation. We compare the performance of the proposed method with the original ViT architecture to validate its effectiveness. In addition, to mitigate data imbalance, we employ the Synthetic Minority Over-sampling Technique (SMOTE) together with the Focal Loss function during training, thereby improving the model's sensitivity and discriminative power

for minority classes. Finally, we investigate the impact of transfer learning on DR grading performance, analyzing how pre-trained weights and domain differences influence model convergence and generalization.

2 Related Work

Manual DR grading places a heavy burden on ophthalmologists, and the rising number of diabetic patients is expected to increase the risk of misdiagnosis. Automated DR grading systems can therefore provide substantial clinical support by enabling earlier detection and reducing subjectivity in decision-making. With advances in computer vision, a variety of deep learning architectures have been explored for this task.

CNNs remain a foundational architecture for medical vision tasks [11–13]. These approaches primarily emphasize data-level refinement or backbone adaptation. Approaches build on these CNN backbones. Li et al. [14] proposed CANet, which uses ResNet-50 as the backbone and incorporates a Disease-Specific Attention Module and a Disease-Dependent Attention Module for joint DR and diabetic macular edema (DME) grading, selectively emphasizing disease-relevant features and inter-disease correlations. Ma et al. [15] proposed an Asymmetric Bi-Classifier Discrepancy Minimization (ABiD) framework, where two classifiers with different decision boundaries are used, and their prediction discrepancy is minimized to improve feature extraction. He et al. [16] proposed CABNet, introducing a Category Attention Block (CAB) and a Global Attention Block (GAB); CAB provides more discriminative regional features for each class, and both modules can be combined with various CNN backbones. Recent studies have explored DR detection using deep learning with a focus on image enhancement and transfer learning. For instance, Abbood et al. [17] developed a hybrid retinal image enhancement algorithm that significantly improves lesion visibility. Jabbar et al. [18] investigated deep transfer learning-based detection, particularly targeting images acquired in resource-limited or remote settings. Lam et al. [19] utilized Contrast Limited Adaptive Histogram Equalization (CLAHE) as a preprocessing step to enhance image contrast, thereby facilitating the grading of DR stages via CNN. Singh et al. [20] introduces a novel deep learning architecture that utilizes a multi-scale residual attention block (MSRAB) and a cross-attention block (CrAB) for DR grading. In contrast, the proposed method focuses on architectural design by introducing a PWAB-guided spatial importance modeling strategy prior to patch embedding, which provides lesion-aware spatial guidance for Transformer-based representation learning. Therefore, our method is complementary to existing enhancement- or transfer-learning-based DR frameworks and can potentially be integrated with such strategies.

Despite the success of CNN-based methods—such as the approach by Singh et al. [20], which utilized CLAHE-based preprocessing to enhance lesion visibility for DR grading—their reliance on local receptive fields often hampers the capture of long-range dependencies and global semantics. This limitation is particularly critical in DR diagnosis, where pathological features may be dispersed across the entire fundus image. To address this, the ViT [10] introduces a paradigm shift by employing multi-head self-attention to model relationships among image patches, effectively capturing global context through positional encodings. This architecture has been further refined through various iterations: SETR [21] and DETR [22] established pure Transformer back-bones for segmentation and detection, while hybrid models like TransUNet [23] and LeViT [24] combine convolutional local feature extraction with Transformer-based global modeling. Specifically for DR grading, recent studies [25–27] have demonstrated that ViT-based models can effectively localize discriminative lesions and consistently outperform traditional CNN architectures across multiple clinical benchmarks. These advancements underscore the potential of self-attention mechanisms to provide more robust and holistically informed diagnostic support compared to localized convolutional filters.

A major challenge in DR datasets is severe class imbalance, where advanced stages (e.g., severe NPDR, PDR) are underrepresented. Nazih et al. [28] addressed this by combining data augmentation with class-balanced weighting, encouraging the model to pay more attention to minority classes. Al-Antary and Arafa [29] combined image augmentation with a multi-level, multi-scale residual encoder and a multi-scale attention mechanism to improve DR severity classification. Beyond direct image-level classification, some work relies on structured representations: Huang et al. [30] proposed RTNet (Relation Transformer Network), which integrates a Global Transformer Block (GTB) and a Relation Transformer Block (RTB) for vessel and lesion segmentation. The resulting lesion maps can then be used for self-supervised DR grading.

In parallel, growing attention has been paid to positional information and long-range dependencies in images, leading to position-aware or position-wise attention mechanisms. Fu et al. [31] introduced a position attention module that aggregates long-range contextual information to strengthen local representations and reduce mis-segmentation. Li et al. [32] proposed an efficient position-aware module optimized for device constraints. Hu et al. [33] designed a local relationship–based feature extractor that quickly guides semantic inference to key regions. Wang et al. [34] proposed a non-local operation that computes responses at a position as a weighted sum of features across all positions, enhancing long-range feature modeling and improving performance in various vision tasks. Overall, position attention and non-local operations are often used to compensate for CNN's locality, enabling joint modeling of local and global features.

Building upon the foundational principles of class imbalance mitigation [28–30] and position-aware long-range modeling [31–34], we propose the Position-Wise Attention Block (PWAB) to enhance patch-level sensitivity and fine-grained lesion representation within the ViT architecture. While previous works primarily utilize position-awareness for global context, our PWAB specifically targets localized feature enhancement at the input stage. This adaptation ensures that the model remains sensitive to subtle clinical indicators—such as microaneurysms—before global dependencies are processed. Furthermore, by incorporating Focal Loss to alleviate the inherent class imbalance of DR datasets, our approach tightly integrates position-sensitive enhancement with the Transformer's global modeling capacity. This strategy effectively addresses both the recognition of subtle lesions and the rarity of advanced DR stages, achieving superior diagnostic performance in a lightweight manner without incurring substantial computational overhead.

In summary, existing attention-based approaches in medical image analysis mainly focus on channel reweighting or global attention modeling within convolutional or Transformer architectures. Although several studies have explored position-aware mechanisms, their implementations and objectives differ from the proposed PWAB. In contrast, our work emphasizes pre-token spatial importance modeling to guide Transformer-based DR grading, which provides a distinct architectural perspective compared with existing methods.

3 Method

This chapter describes the model design and training procedure adopted in this study. We employ the ViT as the backbone and embed a Position-Wise Attention Block within its standard Transformer blocks to enhance patch-level spatial correlations and fine-grained lesion representation, without significantly increasing computational cost. For clarity, we first introduce the basic ViT architecture, then present the motivation and formulation of PWAB, followed by their integration strategy and the loss design for handling class imbalance. This structured exposition provides a complete methodological framework from data representation to model optimization, forming the basis for the subsequent experiments and analyses.

3.1 Vision Transformer Architecture

The overall architecture consists of four main components: Patch Embedding, ViT Encoder, MLP Head, and the final classification layer.

Given an input fundus image, the Patch Embedding module partitions it into non-overlapping patches of fixed size. Specifically, the image of height H and width W is divided into $N = \frac{HW}{P^2}$ patches, where P is the patch resolution. Each patch is then flattened and projected into a trainable feature space, as illustrated in Fig. 1. Next, all patch features are converted into a 1D token sequence suitable for input to the ViT encoder. The sequence is composed of a learnable classification token x_{class} and the flattened vectors of the N patches. The sequence construction procedure is summarized in Eq. (1):

$$z_0 = \left[x_{class}; x_p^1 E; x_p^2 E; \ldots; x_p^N E\right] + E_{pos} \tag{1}$$

where each patch vector is mapped to a fixed dimension D for linear transformation as shown in formula $E \in \mathbb{R}^{(P^2 \cdot C) \times D}$, and a position encoding is added to retain the spatial position information of each patch in the image as shown in formula $E_{pos} \in \mathbb{R}^{(N+1) \times D}$.

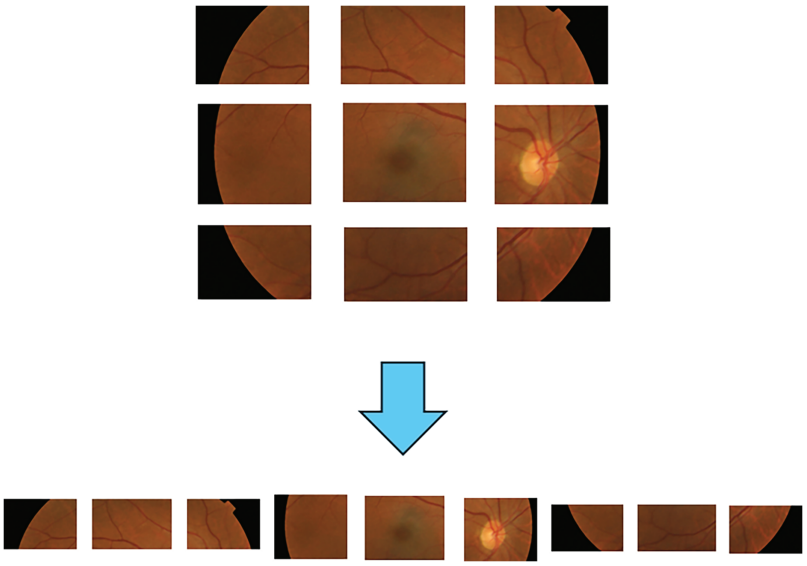

Figure 1: Patches embedding operation diagram.

After patch embedding and sequence construction, sinusoidal or learnable positional embeddings are added to each token to retain spatial order information. The resulting sequence is then fed into the ViT encoder.

The ViT encoder is composed of multiple stacked Transformer blocks. Each block contains two sub-layers: (1) Multi-Head Self-Attention (MHSA) and (2) a Multi-Layer Perceptron (MLP). Layer Normalization (LN) is applied before each sub-layer, and residual connections are applied after each sub-layer to stabilize training and mitigate vanishing gradients.

In the MHSA layer, the normalized input sequence is linearly projected into three sets of vectors: queries Q, keys K, and values V. The attention mechanism then computes pairwise correlations between Q and K and uses these to re-weight V, as defined in Eq. (2). Multiple attention heads are computed in parallel and concatenated, as indicated in Eq. (3), allowing the model to capture diverse relationships in different subspaces.

$$Attention\left(Q,K,V\right)=softmax\left(\frac{QK^{T}}{\sqrt{d_{k}}}\right)V \tag{2}$$

$$MultiHead\left(Q,K,V\right)=\left[head_{i};\ldots;head_{h}\right]W^{O} \tag{3}$$

where $head_i$ denotes the output of the i-th attention mechanism, and W^o represents the learnable weight matrix used to linearly combine and project the concatenated outputs of all h attention heads back into the model's intended dimensionality.

The MLP block in each Transformer layer consists of two fully connected layers separated by a GELU activation, enhancing the non-linear expressiveness of token representations. At the end of the encoder, the class token, which aggregates image-level information, is passed to an MLP Head for DR grade prediction. The MLP Head is structurally similar to the encoder MLP but is dedicated to classification.

When pre-trained ViT weights are used, early embedding and encoder layers are typically frozen or fine-tuned with a smaller learning rate, while the MLP Head is reinitialized and retrained to match the number and characteristics of DR classes. This allows the model to benefit from large-scale pre-training while adapting to the target medical domain.

To further improve ViT's sensitivity to subtle lesions and local spatial cues, we integrate a Position-Wise Attention Block as a lightweight branch parallel to the backbone residual stream. PWAB introduces position-adaptive local enhancement with very small additional parameter and computational cost, and can be seamlessly combined with pre-trained ViT, preserving its global modeling capability.

3.2 Position-Wise Attention Block

Shi et al. [35] proposed a position-wise attention design that strengthens local spatial context modeling in ViT with minimal computational overhead. The core idea is to explicitly estimate contextual correlations between spatial positions using lightweight convolutions and position-wise gain attention, thereby highlighting subtle, scattered DR lesion signals.

DR is a major cause of blindness worldwide, and its lesions are typically small, sparse, and irregularly distributed. Microaneurysms and hard exudates may occupy only a few pixels and appear at varying locations. In addition, fundus images have relatively uniform background colors, and large non-lesion regions can easily distract the model. Under these conditions, relying solely on global features makes it easy to overlook small but clinically crucial localized abnormalities. To address this, we place PWAB in parallel after patch embedding so that neighborhood context and position-dependent gains are computed before the tokens enter the Transformer encoder. This provides a more lesion-aware representation for subsequent global self-attention (Fig. 2).

In PWAB, the input image (or early feature map) first passes through a convolutional layer that adjusts channel dimensionality and enriches semantic content. This step not only increases or reduces dimensionality, but also emphasizes spatial context for subtle lesions, forming a stronger basis for later dependency modeling. Subsequently, three 1×1 convolutional layers are applied for spatial feature extraction. Convolution A produces features that serve as a query-like representation; Convolution B produces key-like features with the same shape; and Convolution C outputs a value-like feature map with 512 channels, encoding deeper semantic information.

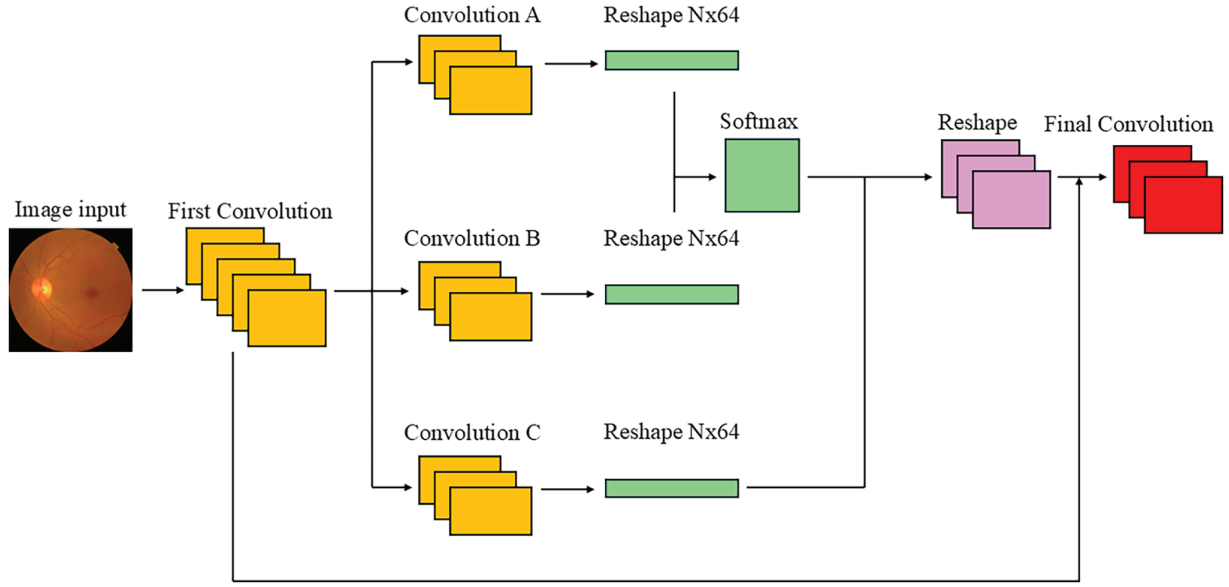

Figure 2: Position-wise attention block architecture.

The Convolution A and Convolution B features are then used to construct a spatial attention map via a softmax-based similarity function, as defined in Eq. (4). In this context, the similarity between vector A_i and B_j is first computed and then processed through the Softmax operator for normalization. The exponential component maps the raw similarity scores to a non-negative domain and amplifies the contrast between scores, enabling the model to assign higher weights to more relevant features. By dividing by the sum of all exponential terms, which is the denominator, the resulting weights form a valid probability distribution that sums to unity:

$$p_{ji} = \frac{\exp(A_i B_j)}{\sum_{i=1}^{N} exp(A_i B_j)} \tag{4}$$

where A and B denote the i and j feature vectors, respectively, and N represents the total number of candidates in the attention pool.

The feature map produced by Convolution C is reshaped and multiplied by the spatial attention map P, producing an output that incorporates spatial correlations across the entire image. Intuitively, this step aggregates context-aware value features according to the learned similarity between positions. The resulting feature map is then reshaped back to the original spatial layout, and a residual connection is applied by adding it to the original feature map that has not undergone spatial attention. The final output of PWAB is given in Eq. (5):

$$O_i = \alpha \sum_{i=1}^{N} \left(P_{ji} C_i \right) + I_j^{'} \tag{5}$$

where C_i denotes the feature vector of the i-th neighboring node, representing the information contributed by the source node in the aggregation step. And I_j is defined as the initial or transformed feature representation of the target node j itself, which is incorporated as a self-connection term to preserve the node's local information during the feature update process.

This design is inspired by the mechanism of Transformer self-attention but emphasizes explicit spatial dependency in the image domain. The initial high-dimensional convolution provides a rich feature basis, enabling broader semantic correlations to be captured even under monotonous color conditions. Convolutions A and B form a "position dependency map" that measures similarity be-tween positions, whereas Convolution C supplies high-dimensional semantic features used for attention weighting. The residual connection avoids over-smoothing and preserves detailed information, while also improving training stability and preventing vanishing gradients. Compared with traditional CNNs, whose fixed receptive fields limit their ability to model long-range dependencies, PWAB learns positional relationships dynamically and can propagate information across distant locations. Compared with full self-attention, PWAB reduces dimensionality through convolutional layers and operates in a more structured manner, substantially lowering the computational cost. Through the combination of convolution and attention, PWAB effectively compensates for ViT's weakness in local feature capture, enhancing fine-grained spatial dependencies while preserving global structural information. This is particularly valuable for DR grading, where small details are crucial.

3.3 Integration of Vision Transformer and Position-Wise Attention Block

The integrated PWAE-ViT architecture is shown in Fig. 3. Although ViT already includes a patch processing module, its main function is to split the input image into fixed-size patches, linearly project them to token vectors, and add positional embeddings before passing them to the Transformer encoder. For DR fundus images, however, subtle lesions such as microaneurysms and hard/soft exudates are scattered, extremely small, and highly sensitive to local contrast. Simply applying global self-attention to patch tokens may not provide sufficient sensitivity to these micro-local cues.

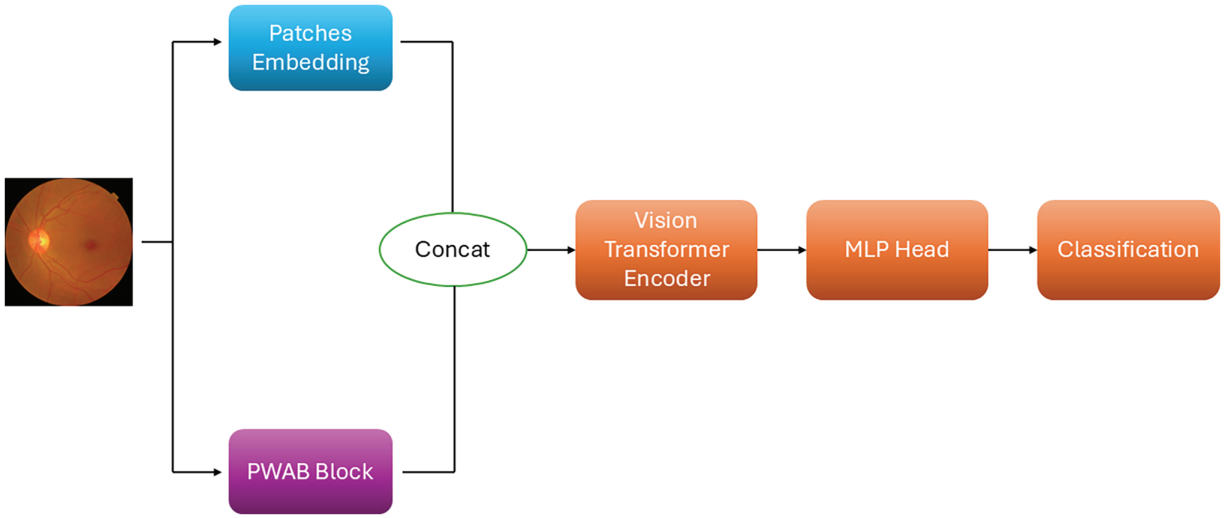

Figure 3: PWAE-ViT model architecture.

To address this, we incorporate PWAB so that, immediately after the image is input, the contextual correlation of each spatial position can be estimated. This design allows the model to consider both global structure and local details simultaneously. The convolutional operators in PWAB strengthen lesion-related regional representations, while the position-wise attention mechanism captures the contextual relationships between "lesion points" and their surroundings. As a result, the model learns a joint representation of global structure and local anomalies.

In the PWAE-ViT framework, the PWAB architecture operates concurrently with the ViT's patch embedding process. As the input image is discretized into a sequence of fixed-length vectors, the PWAB simultaneously computes position-dependent contextual gains at the same input level. Rather than a sequential refinement, the features from both modules are fused—integrating the discretized tokens with localized gain weights—to generate enriched representations. These fused features are subsequently fed into the Multi-Head Self-Attention (MHSA) layers within the Transformer encoder, ensuring that the self-attention mechanism operates on tokens already embedded with fine-grained spatial cues. This fusion yields two key benefits in a complementary and mutually reinforcing manner. First, in terms of representational capability, PWAB enhances local contrast and boundary cues that are critical for detecting small lesions, while ViT preserves strong modeling of long-range and global semantic patterns. By combining these two components, the model can simultaneously attend to fine-grained details and broader structural context, thereby preventing small lesions from being diluted by purely global processing and avoiding over-emphasis on isolated patches without considering overall retinal structure. Second, from the perspective of training stability and scalability, PWAB is integrated as a lightweight residual branch, introducing only a modest increase in parameters and FLOPs. Because the PWAB stream and the native ViT stream learn in parallel and are fused through structured gain operations, feature interference is reduced, and the benefits of ViT pre-training are largely retained. When multiple abnormal patterns (such as hemorrhages and exudates) appear concurrently in an image, the resulting fused representation becomes richer and more discriminative, which in turn enhances the robustness and overall performance of the downstream DR grading task.

3.4 Training Objective and Imbalance Handling

Class imbalance is a major challenge in medical image classification, especially in DR grading. In many public DR datasets, images with No DR are abundant, whereas mild, moderate, severe, and proliferative DR cases are relatively rare. This distribution causes the model to focus primarily on the majority class, often neglecting minority classes, which may lead to clinically serious misclassification of mild or severe DR as normal.

The conventional loss function for classification is cross-entropy (CE). Although CE effectively measures the discrepancy between predicted and true distributions, its performance degrades under severe class imbalance. Majority-class samples dominate the gradients, leading the model to overfit easy majority examples. Minority-class samples contribute relatively little to the loss, resulting in poor recognition of rare classes and a tendency to ignore hard examples, which harms generalization.

To address these issues, we adopt the Focal Loss [36] proposed by Lin et al. as the main training objective. Its core idea is to down-weight the loss of easily classified samples and up-weight hard or minority samples, encouraging the model to focus on ambiguous cases and under-represented classes. The mathematical forms of Focal Loss and cross-entropy are given in Eqs. (6) and (7), respectively:

$$FL\left(p_t\right) = -\alpha\left(1-p_t\right)^r log\left(p_t\right) \tag{6}$$

$$CE\left(p_t\right) = -\sum_{i=1}^{C} y_i log(p_i) \tag{7}$$

In Focal Loss, a class-balancing factor α is introduced to adjust the relative importance of different classes, so that classes with fewer samples receive larger weights during training. In addition, the modulating factor $\left(1-p_t\right)^r$ reduces the contribution of well-classified samples (where p_t is close to 1) and emphasizes those that are misclassified or uncertain. As p_t approaches 1, the term $\left(1-p_t\right)^r$ approaches 0, so its contribution to the loss becomes smaller, and easy samples are prevented from dominating the gradient.

In this study, Focal Loss is further combined with dynamically adjusted class weights, which are updated according to the evolving class-wise performance during training. This strategy enables the model to concentrate on minority or difficult classes, alleviating the impact of long-tailed distributions and improving recall and F1-score for clinically critical but rare DR grades.

4 Experimental Results and Discussion

This section presents the experimental setup and quantitative evaluation of the proposed PWAE-ViT model on the APTOS-2019 and IDRiD Diabetic Retinopathy dataset. We first describe the dataset and implementation details, then introduce the evaluation metrics, followed by overall performance analysis, confusion matrix interpretation, and an ablation study to quantify the contribution of each component.

4.1 Dataset and Experimental Settings

We evaluate our method on the APTOS-2019 and IDRiD Diabetic Retinopathy dataset, which contains 3662 and 516 retinal fundus images labeled with five DR severity grades. The class distribution in APTOS-2019 is as follows: 1805 images with No DR, 370 with mild NPDR, 999 with moderate NPDR, 193 with severe NPDR, and 295 with PDR. And the class distribution in IDRiD is as follows: 168 images with No DR, 25 with mild NPDR, 168 with moderate NPDR, 93 with severe NPDR, and 62 with PDR.

The dataset is randomly split into 80% for training and 20% for validation. Prior to training, all retinal images were resized to a fixed resolution of 224 × 224 pixels. Color normalization was applied to reduce illumination variations across images by normalizing each RGB channel using the dataset mean and standard deviation. No data augmentation was applied during validation or testing. The models are trained for 100 epochs with a batch size of 32, using the Adam optimizer and an initial learning rate of 0.001. All hyperparameters and preprocessing steps were fixed across experiments to ensure fair comparison and reproducibility.

All experiments were conducted on a workstation equipped with an NVIDIA GeForce RTX 4070 Ti Super GPU (16 GB VRAM). The proposed model was implemented using the PyTorch framework (version 2.6.0), with CUDA (version 12.4) acceleration to ensure efficient training and inference.

4.2 Performance Metrics and Classification Performance

To assess classification performance, we report Accuracy, Precision, Recall, and F1-Score [37]. These indices are computed from the confusion matrix using True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Together with the confusion matrix, these metrics reveal not only the overall performance but also the per-class prediction behavior and error distribution.

The proposed PWAE-ViT achieves an overall accuracy of 0.84 on APTOS-2019. The per-class Precision, Recall, and F1-Score are summarized in Table 1. Among the five grades, No DR and moderate NPDR show the most stable performance, with comparatively high precision and recall. In contrast, severe NPDR and PDR exhibit lower recall, which can be attributed to both the limited number of samples and the visual similarity of their lesions to neighboring grades (e.g., moderate vs. severe NPDR, severe NPDR vs. PDR).

Table 1: APTOS-2019 model performance in different classes.

Class	Precision	Recall	F1-Score	Number of Samples
no DR	0.97	0.98	0.98	361
mild NPDR	0.56	0.57	0.56	74

(Continued)

Table 1 (continued)

Class	Precision	Recall	F1-Score	Number of Samples
moderate NPDR	0.68	0.81	0.74	200
severe NPDR	0.55	0.29	0.38	39
PDR	0.61	0.32	0.42	59

Similar trends are observed in the IDRiD dataset, with the results summarized in Table 2. Specifically, the precision for the No DR and Moderate NPDR classes remains the highest among all grades. However, the overall performance on the IDRiD dataset is significantly lower than that on the APTOS-2019 dataset. We attribute this performance gap to the high data requirements of ViT; as noted in [10], ViTs lack the inductive biases inherent in CNNs and typically necessitate large-scale datasets to achieve optimal convergence and generalization.

Table 2: IDRiD model performance in different classes.

Class	Precision	Recall	F1-Score	Number of Samples
no DR	0.90	0.93	0.91	34
mild NPDR	0.24	0.39	0.30	5
moderate NPDR	0.66	0.32	0.43	34
severe NPDR	0.24	0.30	0.26	19
PDR	0.20	0.32	0.24	12

To address the pronounced class imbalance in the datasets, we employ Focal Loss combined with a dynamic class-weighting strategy, shifting the network's focus toward hard, under-represented examples. While CNNs typically excel on smaller datasets due to their inherent inductive biases—such as locality and translation invariance—vanilla ViTs often struggle without large-scale pre-training. Our proposed PWAE-ViT bridges this gap by integrating the PWAB module, which introduces localized feature enhancement at the input level. This architectural improvement compensates for the ViT's lack of inductive bias, allowing the model to capture fine-grained pathological cues effectively even with limited data. As reflected in Tables 3 and 4, our approach not only outperforms the standard ViT but also achieves performance comparable to, or exceeding, state-of-the-art CNNs.

Table 3: Performance comparison of per-class precision on the APTOS-2019 dataset.

Model	No DR	Mild NPDR	Moderate NPDR	Severe NPDR	PDR
CNN	0.97	0.55	0.70	0.55	0.60
ViT	0.87	0.00	0.49	0.00	0.00
PWAE-ViT (Ours)	0.97	0.56	0.68	0.55	0.61

Table 4: Performance comparison of per-class precision on the IDRiD dataset.

Model	No DR	Mild NPDR	Moderate NPDR	Severe NPDR	PDR
CNN	0.65	0.00	0.76	0.61	0.80
ViT	0.48	0.00	0.40	0.00	0.00
PWAE-ViT (Ours)	0.90	0.24	0.66	0.24	0.20

4.3 Confusion Matrix Analysis

To visualize the distribution and direction of classification errors, we analyze the confusion matrix of the proposed model. The diagonal entries correspond to correctly classified samples, while off-diagonal entries indicate misclassifications.

Fig. 4 shows the test results in different classes. It can be found that the sample classification in the no DR and moderate NPDR classes is the most stable, and although other classes also have misclassified samples, they still show good performance for overall accuracy. Conversely, severe NPDR and PDR are affected by scarce samples and feature similarity. The main errors are concentrated in mutual misjudgments with adjacent grades, indicating that the discrimination of borderline cases is still challenging.

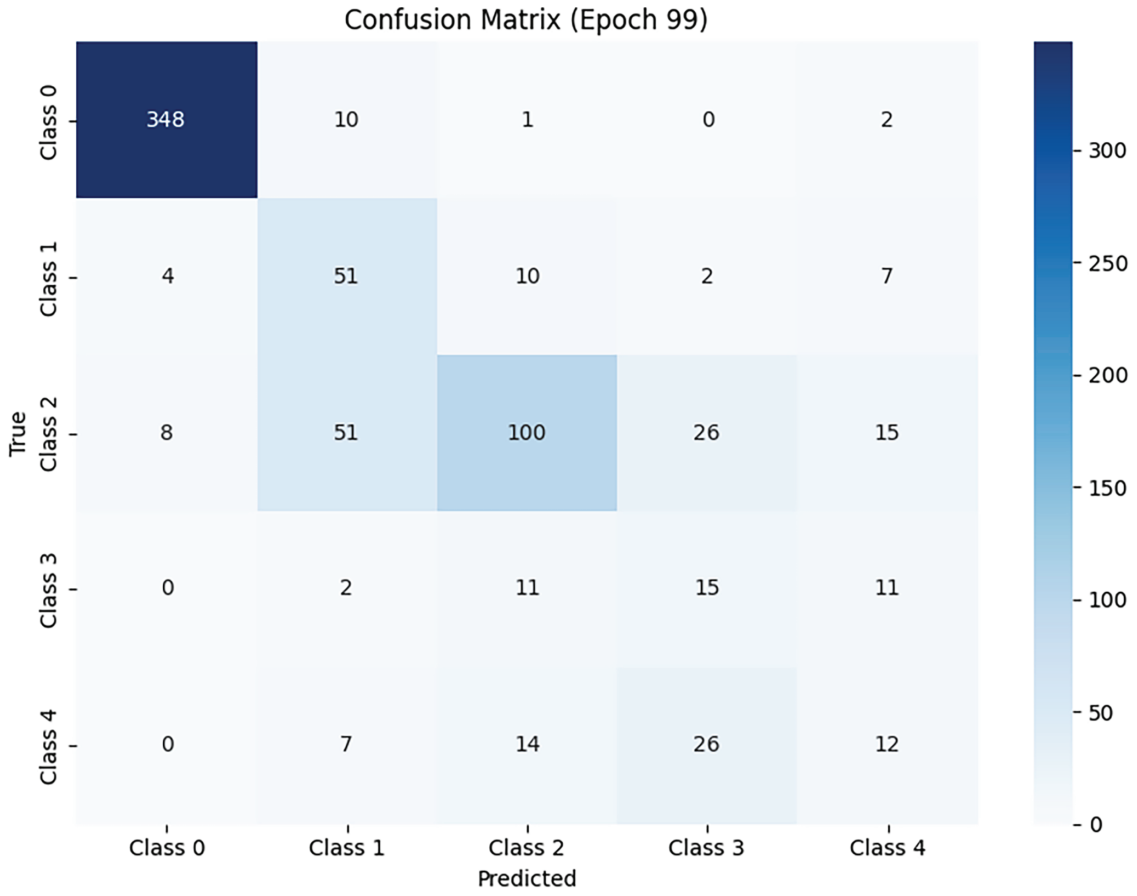

Figure 4: Model performance confusion matrix.

For ease of cross-class comparison, we use a row-wise (per-class) normalized confusion matrix, which can be viewed as the conditional distribution for each true class. This presentation method can more clearly show the recall gap of minority classes. Overall, the confusion matrix, overall accuracy (0.84), and Macro-F1 conclusions mutually confirm that majority classes maintain high precision and recall, while minority classes mainly show conservative misclassification to neighboring grades, rather than random, scattered misclassification.

4.4 Ablation Study

To quantify the contribution of each component, we conduct an ablation study focusing on three factors: (1) the Position-Wise Attention Block, with kernel sizes of 3 × 3, 5 × 5, (2) the use of Focal Loss vs. standard CE Loss, and (3) the effect of ViT pre-trained weights. The results are reported in Table 5.

Table 5: Model ablation study results.

Model	Vision Transformer Pre-Trained Weights	Focal Loss	PWAB	Accuracy
Model 1	V	V	V (3 × 3)	0.84
Model 2	V	V	V (5 × 5)	0.80
Model 3	V	V		0.81
Model4	V		V	0.78
Model 5		V	V	0.61

To further analyze the impact of individual components on different DR severity levels, we additionally report class-wise performance metrics in the ablation study. Specifically, F1-scores for each DR grade are computed to examine which categories benefit most from the proposed modules. The results indicate that the inclusion of PWAB leads to more noticeable improvements in moderate-to-severe DR classes, while maintaining comparable performance for non-DR and mild cases. Detailed comparisons of the F1-scores across various experimental settings are presented in Table 6.

Table 6: Class-wise F1-score comparison on APTOS-2019.

Model	No DR	Mild NPDR	Moderate NPDR	Severe NPDR	PDR
Model 1	0.97	0.55	0.76	0.47	0.46
Model 2	0.97	0.56	0.73	0.37	0.42
Model 3	0.97	0.54	0.75	0.43	0.45
Model 4	0.96	0.46	0.72	0.33	0.38
Model 5	0.82	0.00	0.57	0.00	0.00

Model 1 (full model) includes PWAB, Focal Loss with dynamic weights, and ViT pre-trained parameters. It achieves an accuracy of 0.84, representing the best overall configuration and confirming the effectiveness of the proposed design.

Model 2 removes PWAB while keeping the other settings unchanged, leading to a drop in accuracy to 0.81. This performance degradation shows that PWAB effectively enhances spatial feature modeling across patches. By integrating convolution and attention, PWAB strengthens local feature representation and improves global context understanding, making the architecture more discriminative than the baseline ViT.

Model 3 replaces Focal Loss with standard CE Loss. The accuracy further decreases to 0.78. While this may still be acceptable in balanced image classification tasks, the decline is substantial in the context of DR grading and reflects the strong impact of class imbalance in APTOS-2019. Without Focal Loss, gradients are dominated by "easy" majority-class samples, significantly impairing the recall of rare grades (mild/severe NPDR, PDR) and lowering both overall and macro-level performance.

Model 4 investigates the role of ViT pre-trained weights by training the architecture from scratch solely on APTOS-2019. In this setting, the accuracy drops sharply to 0.61. Moreover, as summarized in Table 7, several minority classes show near-zero precision, indicating that the model fails to learn effective decision boundaries for these grades. Even with PWAB and Focal Loss, the limited dataset size is insufficient to compensate for the lack of prior knowledge.

Table 7: Precision of each class without using ViT pre-trained weights.

	Without ViT Pre-Trained Weights
no DR	75%
mild NPDR	0%
moderate NPDR	48%
severe NPDR	0%
PDR	0%

These findings underscore the importance of transfer learning for data-limited medical imaging tasks. Only by leveraging ViT pre-trained on large-scale datasets, and then incorporating PWAB and imbalance-aware training (Focal Loss + dynamic weights), can the model fully exploit both global modeling capacity and fine-grained lesion sensitivity, thereby achieving the best performance on DR grading.

4.5 Discussion

Using the APTOS-2019 dataset, the proposed PWAE-ViT architecture achieved an overall accuracy of 0.84 (Table 1). With the exception of severe NPDR and PDR, whose performance is constrained by the limited number of samples, the remaining grades exhibit robust precision and recall. In particular, No DR and moderate NPDR show stable behavior, suggesting that the model can reliably distinguish the majority of clinically relevant DR stages.

From a mechanistic standpoint, ViT is well suited for modeling long-range dependencies and global semantics. However, DR lesions are often subtle, scattered, and small (e.g., microaneurysms, hard and soft exudates), making them difficult to capture solely through global self-attention. In the proposed design, PWAB is inserted between the multi-head self-attention and the feed-forward network of each Transformer block. It aggregates local neighborhood context using lightweight operations and then applies position-adaptive gains in a multiplicative residual manner, amplifying fine-grained cues before they are integrated by global attention. In this way, PWAB complements, rather than replaces, the native semantic aggregation capability of ViT, yielding a more balanced local–global representation.

The ablation results (Table 5) further clarify the contribution of each component. Removing PWAB reduces accuracy from 0.84 to 0.81, indicating that position-adaptive modulation provides substantial gains in capturing micro-local structures. When Focal Loss is replaced with standard cross-entropy, accuracy drops to 0.78, highlighting the importance of explicitly addressing the long-tailed distribution: without imbalance-aware training, majority classes dominate the gradients, and the recall of minority grades is markedly degraded. Training the model from scratch without ViT pre-trained weights leads to a much

lower accuracy of 0.61, with several rare classes becoming nearly unrecognizable. Taken together, these findings suggest a clear division of roles: transfer learning supplies a robust and transferable initial feature space, PWAB enhances detail sensitivity on top of this space, and Focal Loss down-weights easy majority samples while emphasizing hard and minority examples. All three are jointly required to achieve clinically usable performance.

From a clinical perspective, the primary value of an automated DR grading system lies in providing consistent, reproducible support for large-scale screening. The PWAB-enhanced ViT maintains high overall accuracy while improving sensitivity to early and subtle lesions, which may help reduce missed diagnoses related to workload, fatigue, or inter-observer variability. The imbalance-aware learning strategy centered on Focal Loss further improves the detectability of advanced but rare grades, an aspect that is particularly important in resource-limited settings where specialist capacity is constrained.

This study nevertheless has several limitations. First, the dataset is relatively small and severely imbalanced, which constrains the achievable performance for severe NPDR and PDR and may exaggerate sensitivity to sampling noise. Second, evaluation is conducted on a single public dataset; as a result, out-of-domain generalization, cross-device robustness, and stability across different acquisition protocols remain to be systematically validated. Third, although PWAB is designed to be lightweight and the number and placement of PWAB layers are unlikely to induce substantial latency, a thorough assessment of computational overhead and deployment trade-offs on real clinical hardware is still required.

In summary, the combination of ViT and PWAB provides a representation that effectively balances global context and local fine-grained details in DR images, a setting characterized by limited data, long-tailed label distributions, and subtle lesion patterns. When integrated with transfer learning and imbalance-aware training, the proposed framework not only yields consistent improvements in quantitative metrics but also enhances the practical feasibility and scalability of automated DR grading toward real-world clinical implementation.

5 Conclusion and Future Work

This study proposed a hybrid architecture that combines a ViT backbone with a Position-Wise Attention Block for DR grading. The central idea is to enhance patch-level local context and fine-grained lesion cues via position-adaptive gains before global encoding, thereby complementing ViT's strength in long-range dependency modeling. On the APTOS-2019 public dataset, the proposed PWAE-ViT outperforms the vanilla ViT in overall accuracy, demonstrating that PWAB can effectively improve sensitivity to subtle and scattered lesions without incurring a substantial increase in parameters or computational cost. In addition, replacing standard cross-entropy with Focal Loss, combined with dynamic class weights, alleviates gradient bias caused by the long-tailed class distribution and enables better focus on clinically critical but rare stages, further enhancing robustness and practical utility.

Looking forward, several research directions are worth pursuing. First, self-supervised and weakly supervised learning represent promising avenues to reduce labeling costs and scale up training data. Reliable annotation of medical images heavily depends on expert effort; thus, learning transferable, general-purpose representations via contrastive pre-training, masked reconstruction, or multi-task prediction, and then achieving accurate DR grading with limited labeled data, will be crucial for cost-effective clinical deployment. Second, the generalizability of the proposed methodology should be validated on a broader spectrum of ophthalmic conditions, such as glaucoma, age-related macular degeneration, and high myopia–related retinopathies, to assess its adaptability across disease types and imaging protocols.

Overall, the PWAE-ViT architecture shows that integrating PWAB with ViT can improve diagnostic accuracy in fundus image classification by explicitly balancing global and local information under conditions of small sample size, long-tailed distributions, and subtle lesion patterns. When combined with transfer learning and imbalance-aware training strategies, the proposed framework provides a solid foundation for developing consistent and scalable automated screening systems, and offers a clear technical starting point for future work toward larger-scale, multi-task, and multi-center clinical applications.

Acknowledgement: The authors would like to thank all individuals and institutions that provided support and assistance for this study.

Funding Statement: This research received no external funding.

Author Contributions: Conceptualization, Yan-Hao Huang; methodology, Yan-Hao Huang; software and experiments, Yu-Tse Huang; data analysis, Yan-Hao Huang; writing—original draft preparation, Yan-Hao Huang and Yu-Tse Huang; writing—review and editing, Yan-Hao Huang; supervision, Yan-Hao Huang. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used in this study are publicly available. APTOS-2019 (https://www.kaggle.com/c/aptos2019-blindness-detection) and IDRiD (https://ieee-dataport.org/open-access/indian-diabetic-retinopathy-image-dataset-idrid) datasets can be accessed from their respective official websites.

Ethics Approval: This study used publicly available, de-identified retinal fundus image datasets. No human participants were directly involved, and therefore ethical approval was not required.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Teo ZL, Tham YC, Yu M, Chee ML, Rim TH, Cheung N, et al. Global prevalence of diabetic retinopathy and projection of burden through 2045 systematic review and meta-analysis. Ophthalmology. 2021;128(11):1580–91. doi:10.1016/j.ophtha.2021.04.027.
2. Savastano MC, Rizzo C, Fossataro C, Bacherini D, Giansanti F, Savastano A, et al. Artificial intelligence in ophthalmology: progress, challenges, and ethical implications. Prog Retin Eye Res. 2025;107:101374. doi:10.1016/j.preteyeres.2025.101374.
3. Wong TY, Sabanayagam C. Strategies to tackle the global burden of diabetic retinopathy: from epidemiology to artificial intelligence. Ophthalmologica. 2020;243(1):9–20. doi:10.1159/000502387.
4. Abramoff MD, Garvin MK, Sonka M. Retinal imaging and image analysis. IEEE Rev Biomed Eng. 2010;3:169–208. doi:10.1109/rbme.2010.2084567.
5. Kadhim AJ, Seyedarabi H, Afrouzian R, Hasan FS. Diabetic retinopathy classification using hybrid color-based CLAHE and blood vessel in deep convolution neural network. IEEE Access. 2024;12:194750–61. doi:10.1109/ACCESS.2024.3519361.
6. Liu T, Chen Y, Shen H, Zhou R, Zhang M, Liu T, et al. A novel diabetic retinopathy detection approach based on deep symmetric convolutional neural network. IEEE Access. 2021;9:160552–8. doi:10.1109/access.2021.3131630.
7. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in neural information processing systems. Vol. 30. Red Hook, NY, USA: Curran Associates Inc.; 2017. doi: 10.65215/ctdc8e75.
8. Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, et al. Transformers in medical imaging: a survey. Med Image Anal. 2023;88(1):102802. doi:10.1016/j.media.2023.102802.
9. Goh JHL, Ang E, Srinivasan S, Lei X, Loh J, Quek TC, et al. Comparative analysis of vision transformers and conventional convolutional neural networks in detecting referable diabetic retinopathy. Ophthalmol Sci. 2024;4(6):100552. doi:10.1016/j.xops.2024.100552.

10. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
11. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. doi:10.1109/CVPR.2016.90.
12. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015 Jun 7–12; Boston, MA, USA. doi:10.1109/cvpr.2015.7298594.
13. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. 2014.
14. Li X, Hu X, Yu L, Zhu L, Fu CW, Heng PA. CANet: cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading. IEEE Trans Med Imag. 2020;39(5):1483–93. doi:10.1109/TMI.2019.2951844.
15. Ma Y, Gu Y, Guo S, Qin X, Wen S, Shi N, et al. Grade-skewed domain adaptation via asymmetric bi-classifier discrepancy minimization for diabetic retinopathy grading. IEEE Trans Med Imaging. 2025;44(3):1115–26. doi:10.1109/tmi.2024.3485064.
16. He A, Li T, Li N, Wang K, Fu H. CABNet: category attention block for imbalanced diabetic retinopathy grading. IEEE Trans Med Imaging. 2021;40(1):143–53. doi:10.1109/tmi.2020.3023463.
17. Abbood SH, Hamed HNA, Rahim MSM, Rehman A, Saba T, Ali Bahaj S. Hybrid retinal image enhancement algorithm for diabetic retinopathy diagnostic using deep learning model. IEEE Access. 2022;10:73079–86. doi:10.1109/ACCESS.2022.3189374.
18. Jabbar A, Naseem S, Li J, Mahmood T, Jabbar MK, Rehman A, et al. Correction: deep transfer learning-based automated diabetic retinopathy detection using retinal fundus images in remote areas. Int J Comput Intell Syst. 2024;17(1):145. doi:10.1007/s44196-024-00557-x.
19. Lam C, Yi D, Guo M, Lindsey T. Automated detection of diabetic retinopathy using deep learning. AMIA Jt Summits Transl Sci Proc. 2018;2017:147–55. doi:10.1109/C2I456876.2022.10051419.
20. Singh AK, Madarapu S, Ari S. Diabetic retinopathy grading based on multi-scale residual network and cross-attention module. Digit Signal Process. 2025;157(1):104888. doi:10.1016/j.dsp.2024.104888.
21. Zheng S, Lu J, Zhao H, Zhu X, Luo Z, Wang Y, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. doi:10.1109/cvpr46437.2021.00681.
22. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: Computer vision—ECCV 2020. Berlin/Heidelberg, Germany: Springer; 2020. p. 213–29. doi:10.1007/978-3-030-58452-8_13.
23. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. TransUNet: transformers make strong encoders for medical image segmentation. arXiv:2102.04306. 2021.
24. Graham B, El-Nouby A, Touvron H, Stock P, Joulin A, Jegou H, et al. LeViT: a vision transformer in ConvNet's clothing for faster inference. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. doi:10.1109/iccv48922.2021.01204.
25. Porwal P, Pachade S, Kokare M, Deshmukh G, Son J, Bae W, et al. IDRiD: diabetic retinopathy-segmentation and grading challenge. Med Image Anal. 2020;59:101561. doi:10.1016/j.media.2019.101561.
26. Oulhadj M, Riffi J, Khodriss C, Mahraz AM, Yahyaouy A, Abdellaoui M, et al. Diabetic retinopathy prediction based on vision transformer and modified capsule network. Comput Biol Med. 2024;175:108523. doi:10.1016/j.compbiomed.2024.108523.
27. Awasthi V, Awasthi N, Kumar H, Singh S, Singh PP, Dixit P, et al. ViT-HHO: optimized vision transformer for diabetic retinopathy detection using Harris Hawk optimization. MethodsX. 2024;13(1):103018. doi:10.1016/j.mex.2024.103018.
28. Nazih W, Aseeri AO, Atallah OY, El-Sappagh S. Vision transformer model for predicting the severity of diabetic retinopathy in fundus photography-based retina images. IEEE Access. 2023;11:117546–61. doi:10.1109/ACCESS.2023.3326528.

29. Al-Antary MT, Arafa Y. Multi-scale attention network for diabetic retinopathy classification. IEEE Access. 2021;9:54190–200. doi:10.1109/ACCESS.2021.3070685.
30. Huang S, Li J, Xiao Y, Shen N, Xu T. RTNet: relation transformer network for diabetic retinopathy multi-lesion segmentation. IEEE Trans Med Imag. 2022;41(6):1596–607. doi:10.1109/TMI.2022.3143833.
31. Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, et al. Dual attention network for scene segmentation. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. doi:10.1109/cvpr.2019.00326.
32. Li X, Zhong Z, Wu J, Yang Y, Lin Z, Liu H. Expectation-maximization attention networks for semantic segmentation. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. doi:10.1109/iccv.2019.00926.
33. Hu H, Zhang Z, Xie Z, Lin S. Local relation networks for image recognition. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. doi:10.1109/iccv.2019.00356.
34. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. doi:10.1109/cvpr.2018.00813.
35. Shi Q, Fan J, Wang Z, Zhang Z. Multimodal channel-wise attention transformer inspired by multisensory integration mechanisms of the brain. Pattern Recognit. 2022;130:108837. doi:10.1016/j.patcog.2022.108837.
36. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. doi:10.1109/iccv.2017.324.
37. Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. Sci Rep. 2024;14(1):6086. doi:10.1038/s41598-024-56706-x.