Like an Ophthalmologist: Dynamic Selection Driven Multi-View Learning for Diabetic Retinopathy Grading

Xiaoling Luo[1], Qihao Xu[2], Huisi Wu[1], Chengliang Liu[3], Zhihui Lai[1], Linlin Shen[1,4,5]*

[1]Computer Vision Institute, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
[2]Shenzhen Key Laboratory of Visual Object Detection and Recognition, Harbin Institute of Technology, Shenzhen, China
[3]Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, China
[4]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen, China
[5]Guangdong Provincial Key Laboratory of Intelligent Information Processing, China
xiaolingluoo@outlook.com, xqh51199597@outlook.com, hswu@szu.edu.cn, liucl1996@163.com, laizhihui@szu.edu.cn, llshen@szu.edu.cn

Abstract

Diabetic retinopathy (DR), with its large patient population, has become a formidable threat to human visual health. In the clinical diagnosis of DR, multi-view fundus images are considered to be more suitable for DR diagnosis because of the wide coverage of the field of view. Therefore, different from the previous single-view DR grading methods, we design a dynamic selection-driven multi-view DR grading method to fit clinical scenarios better. Since lesion information plays a key role in DR diagnosis, previous methods usually boost the model performance by enhancing the lesion feature. However, during the actual diagnosis, ophthalmologists not only focus on the crucial parts, but also exclude irrelevant features to ensure the accuracy of judgment. To this end, we introduce the idea of dynamic selection and design a series of selection mechanisms from fine granularity to coarse granularity. In this work, we first introduce an Ophthalmic Image Reader (OIR) agent to provide the model with pixel-level prompts of suspected lesion areas. Moreover, we design a Multi-View Token Selection Module (MVTSM) that prunes redundant feature tokens and dynamically selects key information. In the final decision stage, we dynamically fuse multi-view features through the novel Multi-View Mixture of Experts Module (MVMoEM), to enhance key views and reduce the impact of conflicting views. Extensive experiments on a large multi-view fundus image dataset with 34,452 images prove that our method performs favorably against state-of-the-art models.

Code — https://github.com/xqh180110910537/SMVDR

Introduction

According to the 10th edition of the Diabetes Atlas published by the International Diabetes Federation (IDF) (Federation 2021), 537 million adults worldwide were living with diabetes in 2021, accounting for roughly one in every ten adults globally. Diabetic Retinopathy (DR) is the most common complication of diabetes and is considered to be one of the leading causes of blindness (Dai et al. 2024). Taking the international DR Grading standard (Wilkinson et al. 2003) as an example, the severity of DR is diagnosed based on a series of DR lesions and can be divided into five levels (grade 0-4): normal, mild, moderate, severe and Proliferative Diabetic Retinopathy (PDR).

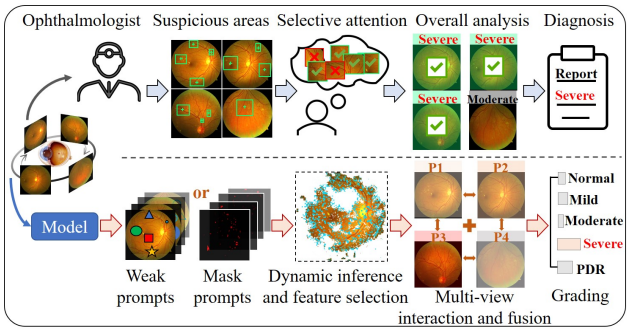


Figure 1: Our approach simulates the process of clinical diagnosis by ophthalmologists, with a coherent diagnostic train of thought from the discovery of suspected areas to the DR grading.

With the development of computer vision (Fang et al. 2024b, 2023, 2024a), researchers have discovered the great potential of artificial intelligence algorithms to work in DR Diagnosis. Although some DR Diagnosis algorithms have been proposed (Dai et al. 2024; Sun et al. 2021), they are still below the level of ophthalmologist diagnosis. According to previous works (Wu et al. 2021), we found that the defects of existing DR Diagnosis methods lie in the following aspects. First of all, most previous works are trained on single-view databases with a field of view (FOV) of only $45°$- $50°$ (e.g. MESSIDOR (Decenciere et al. 2014), EyePACS (EyePACS 2015)). The observable FOV of the human retina can reach $230°$, which means that single-view data is at risk of losing most of the pathological features on the retina. Meanwhile, clinical medical research (Hu et al. 2019) also shows that $45°$single-view mydriatic fundus photography does not meet the technical requirements of DR screening, and fundus multi-view imaging has a better performance in DR clinical diagnosis (Srihatrai and Hlowchitsieng 2018).

*Corresponding Author: Linlin Shen

Secondly, since lesion information plays an important role in DR diagnosis, some works (Sun et al. 2021; Dai et al. 2024) are proposed to improve the performance of the model by lesion-feature enhancement. However, in the actual diagnosis, ophthalmologists tend to ignore the useless parts and focus their attention on the crucial information. The prevailing methods typically enhance lesion features solely to elevate awareness, but do not eliminate redundant features that are irrelevant for diagnostic purposes. Consequently, the model remains susceptible to the distracting influence of such unnecessary information.

To address the above problems, this work proposes an innovative framework to simulate the diagnostic process of ophthalmologists. We adopt the multi-view fundus image data which is more suitable for clinical diagnosis, and design the multi-view DR grading method. Our method can simultaneously learn multi-field fundus images containing complete periocular information. Compared with the single-view DR grading method, the proposed multi-view DR grading method conforms to the actual clinical scenes and can improve the grading accuracy of the model.

Moreover, inspired by the clinical experience of ophthalmologists, we introduce the core idea of dynamic selection into the model. As Fig. 1, when making a diagnosis, ophthalmologists usually mark suspicious lesion areas first, then selectively analyze information, and finally synthesize multiple views for comparative analysis. To simulate the diagnostic ways of ophthalmologists, our model first introduced areas of suspected lesions of fundus images. We used an Ophthalmic Image Reader (OIR) agent to extract the lesion area. This OIR agent has two interfaces, one can generate pixel-level mask prompts of the lesions, and the other can get the weak prompts of the area of the lesions. The prompts are used for further dynamic inference and feature selection. We have observed that when ophthalmologists read fundus images, they focus on suspicious areas, while irrelevant information is excluded. Therefore, we designed a Multi-View Token Selection Module (MVTSM) to analyze each token in the feature map dynamically. MVTSM is capable of selecting valuable tokens and removing useless tokens.

Then, in the comprehensive analysis of multiple views, considering the lesion correlation among multi-view fundus images (i.e., the same lesion could be dispersed in different views), we introduced the Multi-View Interaction Mamba (MVIMamba) block to implement feature interaction between multiple views. Finally, the ophthalmologist selects the diagnosis result of the important view as the final diagnosis. For example, when the patient's DR grade is mild, and the lesion only appears in one view, the ophthalmologist will use the diagnosis result of this view as the final judgment. In this case, the diagnostic results of other views without lesions would interfere with the judgment of the classifier. Therefore, a Multi-View Mixture of Experts Module (MVMoEM) is designed in our model. By promoting the learning of multiple experts, the MVMoE module can obtain the optimal combination of experts for estimating each view, which can realize the dynamic fusion of multi-view features.

Particularly, due to the development of state-space models (Hafner et al. 2020; Gu et al. 2023), especially the Mamba model (Gu and Dao 2023; Zhu et al. 2024), our model backbone adopts the combination of Mamba and Transformer, and improves the scanning direction of sequence modeling of multi-view features. Global and Window-based Mamba (GAWMamba) block and MVIMamba block are introduced in the model, which preserves the local dependency within views and the long-distance spatial relationship between views (Yuan et al. 2021). The proposed model first selects the matrix elements according to the input features, then deduces and selects the feature tokens dynamically, and finally selectively fuses the multi-view features. Combined with the coherent diagnosis idea, this model realizes the feature inference and selection from fine granularity to coarse granularity, which greatly improves the performance of the model.

Compared with the previous methods, this work has the following contributions:

- We propose an innovative multi-view DR grading method to simulate ophthalmologists, which is combined with coherent diagnosis ideas to achieve feature inference and selection from fine- to coarse-grained features. The model can gradually select the key information and discard the useless interference features to improve the performance of the model.
- In the early stage of fundus-image reading, we introduced the OIR agent to provide the model with prompts of the lesion areas. These prompts are beneficial for subsequent feature inference and selection.
- As for feature analysis, we design the MVTSM that can recognize the redundant feature tokens and prune the redundant feature tokens, to obtain the discriminative features.
- To realize the comprehensive diagnosis, the proposed MVMoEM can dynamically adjust the multi-view fusion strategy to deepen the impression of crucial views. Then, the effect of noise disturbance is reduced, and the robustness of the multi-view DR grading model is improved.

Related Work

End-to-End Learning in DR Grading

Rrecently, deep learning algorithms (Liu et al. 2022a, 2024a) are widely used in DR grading task. Pao et al. (Pao et al. 2020) introduced an innovative bi-channel CNN architecture, which intelligently integrates entropy image grayscale and green-channel features. To address the shortcomings of CNNS in long-distance feature learning, Wu et al. (Wu et al. 2021) proposed a DR grading model based on Transformer(Dosovitskiy et al. 2021). Then, Huang et al. (Huang et al. 2024) combined self-supervised learning with the Transformer block for pre-training to improve the learning ability of the model. However, current DR grading methods predominantly rely on single-view fundus images, typically offering a limited 45°-FOV centered on the macula, neglecting the lesions present in other-filed fundus images. To address this limitation, we propose an innovative multi-view DR grading method. This work integrates features from multiple perspectives, enabling it to capture the most crucial diagnostic areas, including the posterior polar region and peripheral fundus, thereby enhancing diagnostic accuracy.

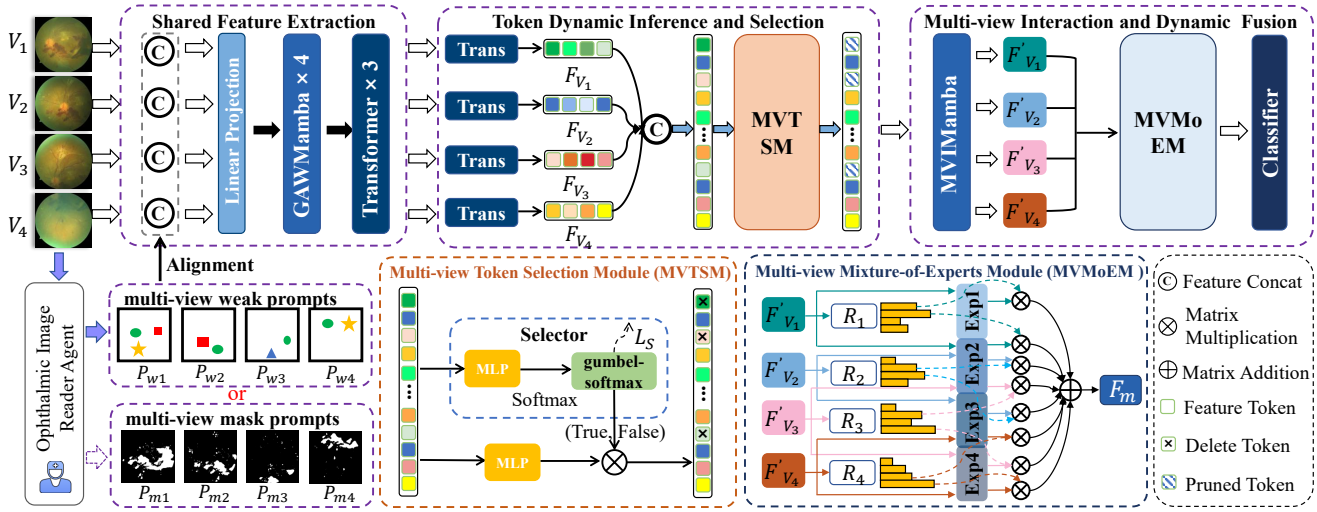


Figure 2: The framework diagram of our proposed method. This method is mainly divided into three stages. The first stage is to extract prompts-enhanced features. The second stage is token dynamic inference and selection implemented by MVTSM, and the third stage is realizing multi-view interaction and dynamic fusion by MVIMamba block and MVMoEM.

Lesion-Fusion Learning in DR Grading

Sun et al. (Sun et al. 2021) proposed an innovative Lesion Aware Transformer model (LAT) for DR Grade prediction and lesion discovery. Further, Luo et al. (Luo et al. 2024) improved the sensitivity of the network to the lesion by combining the lesion snapshot, thus improving the accuracy of DR Classification. Dai et al. (Dai et al. 2024) proposed a framework that includes the image quality assessment subnetwork, focal awareness subnetwork, and DR grading subnetwork. The lesion features extracted by the lesion awareness subnetwork are fascinated with those extracted by the DR grading subnetwork to improve diagnostic accuracy. Since not all areas in the input image are equally important, enhancing lesion information can improve the performance of the model. But actually, in addition to enhancing useful information, pruning redundant information or noise is equally important for model learning. Therefore, we introduce the idea of dynamic selection and design a series of feature selection learning schemes for both fine granularity and coarse granularity.

Methodology

This section describes the main components of our proposed selection-driven multi-view DR grading method (SMVDR), beginning with the addition of multi-view prompts produced by the OIR agent to enhance lesion learning. OIR agent can be composed of any prompt generation model. Here, we use HACDR-Net (Xu et al. 2024) as the prompt generator to generate mask prompts (MPs), and randomly select part of the region from the mask to form weak prompts (WPs). OIR agent generates four types of lesion masks including hard exudates (EX), soft exudates (SE), microaneurysms (MA), and hemorrhages (HE). To cope with different scenarios, we utilize WPs and MPs respectively for feature enhancement, where our model using WPs is named SMVDR-W, and the model using MPs is named SMVDR-M. Subsequently, the enhanced features are selectively extracted by the Mamba-Transformer backbone, and the features are dynamically deduced and selected by MVTSM. The final part is multi-view feature interaction and dynamic fusion through MVIMamba block and MVMoEM.

Mamba-Transformer Backbone

An overview of our proposed method can be seen in Fig. 2. The Mamba-Transformer backbone has a hierarchical architecture consisting of 4 GAWMamba blocks, 3 Transformer blocks with self-attention mechanisms (Dosovitskiy et al. 2021), and some functional modules. In this section, we first revisit the mamba preliminaries, and then present the detailed design of the proposed GAWMamba block.

Mamba Preliminaries. Inspired by Vision Mamba (Vim) (Zhu et al. 2024), the State-Space Models (SSMs) (Gu and Dao 2023) applied to discrete image features can convert discrete 2D images into 1D inputs for discrete state-space model equations. Vim transforms a 1D discrete input $x_t \in \mathbb{R}$ to $y_t \in \mathbb{R}$ via a learnable hidden state $h_t \in \mathbb{R}^{\hat{N}}$ with discrete parameters $\bar{A} \in \mathbb{R}^{\hat{N}\times\hat{N}}$, $\bar{B} \in \mathbb{R}^{1\times\hat{N}}$, and $\bar{C} \in \mathbb{R}^{1\times\hat{N}}$ as follows:

$$\begin{aligned} h_t &= \bar{A}h_{t-1} + \bar{B}x_t, \;\; y_t = Ch_t, \;\; \bar{A} = e^{\Delta A}, \\ \bar{B} &= (\Delta A)^{-1}(e^{\Delta A} - I)\cdot \Delta B, \;\; \bar{C} = C. \end{aligned} \tag{1}$$

According to the zero-order hold (ZOH) (Karafyllis and Krstic 2011), $\bar{A}$ and $\bar{B}$ are continuous A and B converted to discrete evolution parameters using a timescale parameter Δ. $\bar{C}$ represents the projection parameters. In addition, the models compute output through a global convolution as in the following:

$$y = x * \bar{K}, \;\; \bar{K} = (\bar{C}\bar{B}, \bar{C}\bar{A}\bar{B}, \ldots, \bar{C}\bar{A}^{\hat{N}-1}\bar{B}), \tag{2}$$

where $\hat{N}$ is the length of the 1D input x, and $\bar{K}$ is a structured convolutional kernel.

GAWMamba Block. The structure of GAWMamba is shown in Fig. 3(a). GAWMamba introduces a novel Global-Window State-Space Model (GWSSM) with a global-window selective scanning mechanism, which can achieve selective extraction of long-distance and local spatial features. We first divide the 2D image from each view into 1D $p * p$ size patches as tokens and add positional embeddings to image tokens and the new class token. Then, we input the features from each view F_v into the GWSSM and Feed-forward Network (FFN) structures, interleaved with Layer Normalization (LN) and residual operations. The output $\tilde{F}_v$ is calculated as follows:

$$F_v = \text{GWSSM}(\text{LN}(F_v)) + F_v, \tag{3}$$

$$\tilde{F}_v = \text{FFN}(\text{LN}(F_v)) + F_v. \tag{4}$$

In the GWSSM, the global-window selective scanning mechanism ensures that while selectively extracting global features through Mamba, it also enhances the ability to capture local details in different regions of the fundus image. As shown in Fig. 3(a), vertical selective scan (VSScan) and horizontal selective scan (HSScan) operations are employed to extract global features of the image. The window selective scan (WSScan) operation independently learns the features within each $k \times k$ window divided in the image, and then reconstructs the complete image features. These scanning operations are all implemented by performing positional transformations and reconstructions on the input tokens. All transformed tokens are scanned using Eq. (1) to obtain the new features. The overall GWSSM can be expressed as:

$$z = \text{SiLU}(\text{Linear}(x)), \tag{5}$$

$$\hat{x} = \text{SiLU}(\text{Conv1d}((\text{Linear}(x)), \tag{6}$$

$$\hat{y} = \text{LN}(\text{HSScan}(\hat{x}) + \text{VSScan}(\hat{x}) + \text{WSScan}(\hat{x})), \tag{7}$$

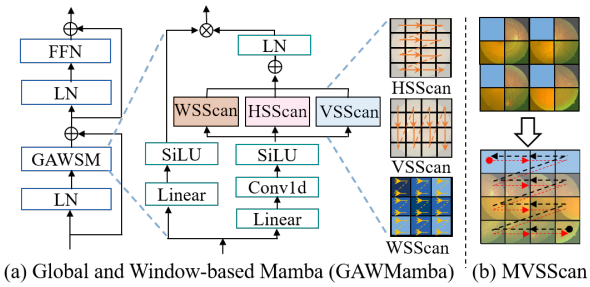
$$\tilde{y} = \hat{y} \odot z, \tag{8}$$

where x represents the input, $\tilde{y}$ represents the output, and operation $\odot$ denotes the Hadamard product. The input x passes through a fully connected layer, a 1D convolution, and the SiLU activation. It is then scanned through multiple branches according to Eq. (1). The final normalized results are gated by z and added together to produce the output $\tilde{y}$.

Multi-View Token Dynamic Selection

By observing the clinical diagnosis process, we find that the ophthalmologist usually only needs to focus on a few key sites to arrive at a diagnosis. Redundant areas and information can reduce the accuracy of multi-view classification. To dynamically select key regions across multiple views like an ophthalmologist, inspired by pruning mechanisms (Liu et al. 2024b,c), we propose the MVTSM that employs confidence-based gated perception to identify areas of interest.

The features synthesized with tokens from N views $\{F_{v_n}\}, n \in \{1, 2, ..., N\}$, are as the input of the confidence-aware function $\mathcal{D}_c$ in MVTSM. The confidence-aware function $\mathcal{D}_c$ consists of MLPs and a softmax function. This function uses the MLP to gradually reduce the token channels to 2, thereby obtaining the confidence c_i for each token in F_s.



(a) Global and Window-based Mamba (GAWMamba) (b) MVSScan

Figure 3: Left: (a) Structure of the GAWMamba block. Right: (b) The scan way of MVSScan in MVIMamba block.

In the phase of inference, the token selection is determined by applying argmax to the confidence scores. For training, it employs Gumbel-softmax (Jang, Gu, and Poole 2017) to enable the differentiation of discrete decisions in Eq. (11). This approach allows for the dynamic selection of multi-view tokens through gating. Then, employing gating operations on the selection results $S = [S_1, S_2, ..., S_i, ..., S_M]$ to obtain the final selected features $\widetilde{F}_s$. The overall process can be expressed as:

$$F_s = \text{Concat}([F_{v_1}, F_{v_2}, ..., F_{v_N}]), \tag{9}$$

$$c_i = \mathcal{D}_c(F_s) = \text{Softmax}(\text{MLP}(F_s)), \tag{10}$$

$$S_i = \frac{\exp\left(\frac{\log(c_i)+g_i}{\tau}\right)}{\sum_{j=1}^{M} \exp\left(\frac{\log(c_j)+g_j}{\tau}\right)} \in \{0,1\}^M, \tag{11}$$

$$\widetilde{F}_s = \text{MLP}(F_s) \odot S, \tag{12}$$

where F_s represent the spliced multi-view tokens, N represents the number of views, M denotes the token count, S_i signifies the i-th token selector mask, c_i is the i-th token selector confidence, and g_i is the i-th token Gumbel noise. Here, $\tau = 1$ is the temperature parameter used to smooth the output.

To dynamically learn the selection rates of multi-view tokens, we introduce a regularization selection loss function $\mathcal{L}_s$. This function computes the mean square error between the actual selection rate and the target selection rate t. The $\mathcal{L}_s$ is computed as:

$$\mathcal{L}_s = (\sum_{i=1}^{M} S_i/M - t)^2. \tag{13}$$

Multi-View Interaction and Fusion

This section specifically describes the MVIMamba block and MVMoEM for processing multi-view features.

MVIMamba Block. To selectively interact with multi-view features, we propose MVIMamba, which is based on SSM and introduces a Multi-view Selection Scan (MVSScan) operation, allowing key multi-view features to fully interact.

As shown in Fig. 3(b), the multi-view tokens $\widetilde{F}_s$ from MVTSM are aligned by position, and then the tokens at each position concatenate in the order of the views. Apart from the scan way, the transformed tokens F'_m use a structure similar to GASSM to interact with multi-view features. It is worth mentioning that MVSScan adopts a bidirectional selective scanning approach. In other words, the scanning view order is 1st to N-th and N-th to 1st. MVSScan performs interactive learning on the selected multi-view features, enhancing its representation ability. Subsequently, the multi-view tokens are restored by position.

MVMoEM. In the final decision-making stage, we introduce MVMoEM to enhance key views and mitigate the impact of conflicting ones. As illustrated in Fig. 2, this module employs a routing mechanism (Riquelme et al. 2021; Zhou et al. 2022) for shared experts across multiple views.

We assign an independent router to each view feature, and these view features are routed to a shared pool of experts based on the router selection. Specifically, MVMoEM assigns four independent routers R_1-R_4, for the four views. Each router R_i comprises a sparse gating network $G(\cdot)$. Finally, a weighted sum of the perspectives from all views F_m is computed in Eq. (14). The routing design of shared experts not only strengthens the interactive decision-making capabilities across views, but also assigns different levels of importance to each view using top-k gating weights. This reduces the impact of view conflicts on the final decision.

$$F_m = \sum_{i=1}^{N}\sum_{j=1}^{e} G(F^{'}_{v_i})_i * E_j(F^{'}_{v_i}) \tag{14}$$
$$= \sum_{i=1}^{N}\sum_{j=1}^{e} TopK(\frac{\exp(W * F'_{v_i} + b)_j}{\sum_{k=1}^{e}\exp(W * F'_{v_i} + b))_k}) * E_j(F^{'}_{v_i}),$$

where N is the number of views, e represents the number of experts, and $E(F'_{v_i})_j$ denotes the j-th expert (implemented as a Feed-Forward Network (FFN)). And we use a Top-2 sparse gating mechanism.

To ensure the balance and stability of the multi-view mixture of experts, an entropy regularization loss function is used. Let P_i^j represent the probability value of the p-th expert in the i-th view. First, the entropy of the expert attention $\mathcal{H}(P_i^j)$ for each view is computed. Compute their average and sum. Then, the overall average expert attention entropy $\mathcal{H}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{e} P_i^j\right)$ is calculated, and $\mathcal{L}_f$ is applied to these values in Eq. (15) and (16).

$$\mathcal{L}_f = \left| \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{e}\mathcal{H}(P_i^j) - \mathcal{H}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{e} P_i^j\right)\right|, \tag{15}$$

$$P_i^j = \frac{\exp(W * F'_{v_i} + b))_j}{\sum_{k=1}^{e}\exp(W * F'_{v_i} + b))_k}. \tag{16}$$

Loss Function. The overall loss $\mathcal{L}_{total}$ for our framework can be defined as:

$$\mathcal{L}_{total} = \mathcal{L}_c + \alpha\mathcal{L}_s + \beta\mathcal{L}_f, \tag{17}$$

Method	Acc.	Spe.	Kappa	F_1	AUC
Inception_ResNet _V2_BSV	70.66	67.13	38.69	65.43	85.55
MobileNet_V2_BSV	72.38	68.73	43.61	67.28	86.88
ResNet50_BSV	73.22	73.20	45.29	69.35	86.76
ResNext50_ 32x4d_BSV	73.36	73.04	47.15	70.37	88.22
ConvNeXt-B_BSV	75.96	77.81	53.72	73.65	89.77
Swin-B_BSV	75.08	75.53	51.32	72.42	88.83
Vim_BSV	77.03	81.20	56.31	75.34	90.67
RETFound_BSV	71.78	70.96	43.67	67.37	86.10
SMVDR-W (Ours)	<u>83.03</u>	<u>88.52</u>	<u>68.89</u>	<u>82.42</u>	<u>95.16</u>
SMVDR-M (Ours)	**84.01**	**91.30**	**71.36**	**83.69**	**95.58**

Table 1: Comparison of the single-view methods and our proposed models. The best results are highlighted in bold, and the sub-optimal results are underlined. (Unit: %)

where $\mathcal{L}_c$ employs the focal loss (Lin et al. 2017), and α and β are two hyper-parameters. $\mathcal{L}_s$ and $\mathcal{L}_f$ have already been introduced in Eq. (13) and Eq. (16).

Experiments

Experimental Setups

Dataset. We conducted experiments on a public large-scale multi-view DR dataset (i.e. MFIDDR [1]). The dataset contains 34,452 color fundus images in JPG format, which were taken by the Zeiss Visucam NM/FA camera. Each eye sample has images of four fields of view (i.e., V1-V4), V1 focused on the macula, V2 centered on the optic disc, and V3-V4 tangent to the upper and lower horizontal lines of the optic disc respectively. All samples are labeled with DR grades according to international standards. The assigned training set contained 25,848 images and the test set contained 8,604 images.

Implementation Details. All methods are implemented by PyTorch, and we conduct all experiments on a single NVIDIA GTX 3090 GPU with 24GB of memory. The batch size and training epoch are set to 8 and 100. We use the Adam optimizer with a learning rate of 0.00001 and dynamically adjust it using a cosine annealing scheduler.

Compared Methods. This experiment uses several open-source methods, which are roughly divided into deep learning models and foundation models: Inception_Resnet_V2 (Szegedy et al. 2016), MobileNet_V2 (Sandler et al. 2018), ResNet50 (He et al. 2016), ResNext50_32x4d (Xie et al. 2017), Swin-B (Liu et al. 2021), ConvNeXt-B (Liu et al. 2022b), Vim (Zhu et al. 2024), MVCINN (Luo et al. 2023), ETMC (Han et al. 2022), LFMVDR (Luo et al. 2024), and RETFound (Zhou et al. 2023).

Evaluation Metrics. To verify the model, we used the common evaluation metrics (Trevethan 2017), such as accuracy (Acc.), precision (Prec.), sensitivity (Sens.), specificity (Spec.), and AUC values (Davis and Goadrich 2006). Notably, to avoid problems caused by sample imbalance, we

[1] https://github.com/mfiddr/MFIDDR

Method	Grade 0			Grade 1			Grade 2			Grade 3			Grade 4		
	Prec.	Sens.	F_1	Prec.	Sens.	F_1	Prec.	Sens.	F_1	Prec.	Sens.	F_1	Prec.	Sens.	F_1
Inception_ ResNet_V2_BSV	74.83	**97.15**	84.54	46.28	12.53	19.72	48.98	39.34	43.64	63.38	60.81	62.07	66.67	15.38	25.00
MobileNet_V2_BSV	76.60	96.70	85.49	48.55	14.99	22.91	56.85	45.36	50.46	64.16	75.00	69.16	60.00	15.38	24.49
ResNet50_BSV	79.38	94.68	86.36	48.34	22.82	31.00	51.48	47.54	49.43	62.94	72.30	67.30	70.00	17.95	28.57
ResNext50_ 32x4d_BSV	79.51	95.13	86.62	46.64	27.96	34.97	55.04	38.80	45.51	71.23	70.27	70.75	75.00	23.08	35.29
ConvNeXt-B_BSV	82.62	95.13	88.43	57.14	34.00	42.64	55.83	49.73	52.60	64.57	76.35	69.97	81.82	23.08	36.00
Swin-B_BSV	81.03	95.43	87.64	54.36	29.31	38.08	54.02	51.37	52.66	68.87	70.27	69.57	92.86	**33.33**	**49.06**
Vim_BSV	85.34	95.13	89.97	60.89	48.77	54.16	45.68	40.44	42.90	66.20	63.51	64.83	98.90	5.13	9.76
RETFound_BSV	77.90	95.65	85.87	49.65	15.66	23.81	44.44	52.46	48.12	64.54	61.49	62.98	73.33	28.21	40.74
SMVDR-W (Ours)	91.09	94.98	92.99	69.11	67.56	68.33	**60.65**	51.37	55.62	**71.52**	76.35	73.86	99.91	25.64	40.82
SMVDR-M (Ours)	**93.48**	93.55	**93.52**	**71.15**	**72.26**	**71.70**	60.00	**60.66**	**60.33**	69.41	**79.73**	**74.21**	**99.99**	17.95	30.43

Table 2: Comparison of precision (Prec.), sensitivity (Sens.), and F_1 scores between our method and single-view methods in DR 0-4 grades. The best results are highlighted in bold, and the sub-optimal results are underlined. (Unit: %)

also introduced kappa and F_1 scores (Sasaki 2007) for comprehensive evaluation of the model.

Experimental Results

Comparison with Single-View Methods. Most of the previous methods are based on single-view fundus images, so to verify the performance of our multi-view DR grading method, we compare our method with single-view methods. Our method uses data from V1-V4 views, while the single view method separately uses four views as training data. We select the view with the highest accuracy among the four-view results of each single-view method, as the comparison results. The experimental results are shown in Table 1. We add the best single-view result identifier _BSV after the single-view method name to distinguish it. It can be seen from the experimental results that our proposed SMVDR-W and SMVDR-M obtained an accuracy of 83.03% and 84.01% respectively, which outperform the accuracy of single-view models by 4.69%-13.35%. The sensitivity, kappa, F_1 score, and AUC value of our models also achieved the top two results.

Further, we compare the classification performance of the competing methods by comparing the precision, sensitivity, and F_1 score of each DR grade. As shown in Table 2, our method achieves the best precision results in each grade, indicating that our proposed SMVDR-W and SMVDR-M are very discriminative for each DR grade. In the results of sensitivity and F_1 scores, although some results are not the best, the performance is still not bad. According to the experimental results, the effect of DR grading using multi-view models is much better than that using single-view models.

Comparison with Multi-View Methods. To better evaluate our method, we compare some existing multi-view methods and introduce the foundation model RETFound (Zhou et al. 2023) as the comparison method. Since RETFound is based on single-view fundus images data, we design a multi-channel network with RETFound as the backbone, and concat the extracted multi-view features as the input of the classifier. This design makes RETFound adapt to multi-view fundus image data. As shown in Table 3, the most effective method is the SMVDR-M model using mask prompts, followed by the SMVDR-W model using weak prompts. Our SMVDR-M model achieves the best performance with average improvements of 4.55%, 9.41%, 10.70%, 6.04% and 3.47% in accuracy, specificity, kappa, F_1 score, and AUC value. Moreover, as shown in Table 4, the proposed SMVDR-M model and SMVDR-W model achieve optimal or sub-optimal results on most of the metrics tested for each DR grade. Overall, the proposed method gets better results in comparison with the other multi-view methods, which further proves the effectiveness of our method.

Method	Acc.	Spe.	Kappa	F_1	AUC
RETFound	74.06	73.83	48.44	70.91	89.04
MVCINN	80.10	83.32	62.45	78.86	91.07
ETMC	81.54	83.44	64.76	79.74	93.53
LFMVDR	82.15	86.97	66.99	81.26	94.81
SMVDR-W (Ours)	83.03	88.52	68.89	82.42	95.16
SMVDR-M (Ours)	**84.01**	**91.30**	**71.36**	**83.69**	**95.58**

Table 3: Comparison of multi-view methods and our proposed models. The best results are highlighted in bold, and the sub-optimal results are underlined. (Unit: %)

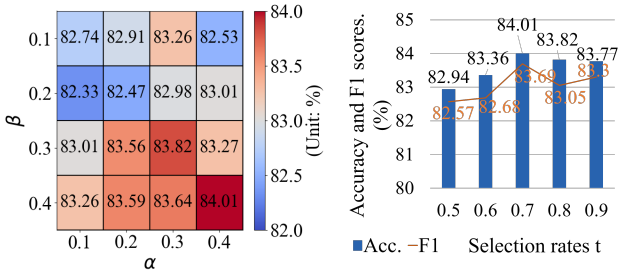
Ablation Studies

To ensure the validity of our model design, we conduct experiments to assess the contributions of each component in the proposed model. Ablation studies are conducted based on the SMVDR-M model, and the experimental results are shown in Table 5.

Analysis for Mamba-Transformer Backbone. Our model uses a Mamba-Transformer backbone to improve the modeling capability of long-distance spatial dependencies. According to lines 1,2, and 5 of Table 5, the results of the backbone network only using Mamba blocks or Transformer blocks are not as good as those using combined blocks. As known in the analysis of 5 on lines 3, 4, and 5, combining global feature learners (e.g., $\mathcal{T}_\phi$ and $\mathcal{M}_\alpha$) and local feature learners ($\mathcal{M}_\beta$) can promote learning ability of the model.

Method	Grade 0			Grade 1			Grade 2			Grade 3			Grade 4		
	Prec.	Sens.	F1	Prec.	Sens.	F_1	Prec.	Sens.	F_1	Prec.	Sens.	F_1	Prec.	Sens.	F_1
RETFound	80.11	96.33	87.47	50.20	27.96	35.92	54.41	40.44	46.39	65.79	67.57	66.67	90.00	23.08	36.73
MVCINN	86.71	96.33	91.26	68.25	48.10	56.43	57.44	**61.20**	59.26	70.00	66.22	68.06	68.42	33.33	44.83
ETMC	86.79	**97.53**	91.85	73.26	56.38	63.72	**66.41**	47.54	55.41	64.41	77.03	70.15	0.12	**90.15**	0.87
LFMVDR	89.69	95.20	92.36	69.53	63.31	66.28	62.05	56.28	59.03	69.48	72.30	70.86	54.55	61.54	**57.83**
SMVDR-W (Ours)	91.09	94.98	92.99	69.11	67.56	68.33	60.65	51.37	55.62	**71.52**	76.35	73.86	99.91	25.64	40.82
SMVDR-M (Ours)	**93.48**	93.55	**93.52**	**71.15**	**72.26**	**71.70**	60.00	60.66	**60.33**	69.41	**79.73**	**74.21**	**99.99**	17.95	30.43

Table 4: Comparison of precision (Prec.), sensitivity (Sens.), and F_1 scores between our method and multi-view methods in DR 0-4 grades. The best results are highlighted in bold, and the sub-optimal results are underlined. (Unit: %)



(a) (b)

Figure 4: Evaluation of the hyperparameters. Comparative analysis of (a) the joint accuracy (unit: %) with parameters α and β in the loss function, and (b) the selection rates t.

Impact of Adding Prompts. Lesion prompts from the OIR agent are added to improve the effect of feature selection. To prove that prompts work, we remove the weak prompts and mask prompts from the overall model. As shown in rows 5 and 6 in Table 5, the accuracy of the model that removed the lesion prompts decreased by 2.37%, which indicates that the addition of the prompts plays an important role in model learning.

Effectiveness of MVTSM. To verify the effect of the dynamic selection of feature tokens on the model, we remove MVTSM from the overall model. As can be seen from rows 5 and 7 in Table 5, MVTSM has a positive effect on improving model performance.

Effectiveness of MVIMamba. Considering the correlation of features among multiple views, this method uses the MVIMamba block to carry out multi-view feature interaction before multi-view fusion. As shown in the results of rows 5 and 8 in Table 5, we verify the validity of MVIMamba for the overall model by removing it.

Effectiveness of MVMoEM. To realize the effective fusion of multiple views and the comprehensive diagnosis of DR, MVMoEM is used to improve the influence of important views and reduce the interference and conflict of redundant views. The results of rows 5 and 9 in Table 5 demonstrate that the multi-view dynamic fusion realized by MVMOEM, has a positive effect on DR grading.

Method	Acc.	Prec.	Spec.	kappa	F_1
$\mathcal{T}_\phi$	83.17	83.16	89.89	69.55	82.77
$\mathcal{M}_\alpha + \mathcal{M}_\beta$	83.22	83.19	89.05	69.35	82.36
$\mathcal{M}_\alpha + \mathcal{T}_\phi$	81.96	82.99	91.29	67.93	81.78
$\mathcal{M}_\beta + \mathcal{T}_\phi$	83.40	83.52	90.37	70.16	82.98
$\mathcal{M}_\alpha + \mathcal{M}_\beta + \mathcal{T}_\phi$	**84.01**	**84.45**	**91.30**	**71.36**	**83.69**
$w/o\ prompts$	81.64	82.11	89.92	66.94	81.38
$w/o\ MVTSM$	83.68	83.99	90.68	70.65	83.39
$w/o\ MVIMamba$	83.87	83.71	90.34	70.84	83.39
$w/o\ MVMoEM$	83.06	81.96	88.93	65.99	82.09

Table 5: Results of Ablation Studies. The overall model is denoted as '$\mathcal{M}_\alpha + \mathcal{M}_\beta + \mathcal{T}_\phi$', where '$\mathcal{M}_\alpha$', '$\mathcal{M}_\beta$', and '$\mathcal{T}_\phi$' indicate the backbone components: mamba block with the global scanners, mamba block with the window scanners, and Transformer block. '$w/o\ prompts$', '$w/o\ MVTSM$', '$w/o\ MVIMamba$' and '$w/o\ MVMoEM$' represent removing prompts, MVTSM, MVIMamba, and MVMoEM from the overall model, respectively. The best results are marked in bold. (Unit: %)

Hyperparameter Sensitivity. We conduct hyperparameter evaluation on our SMVDR-M model. The parameters α and β in the loss function $\mathcal{L}_{total}$ are selected from 0.1, 0.2, 0.3, 0.4. The joint results are demonstrated in the heatmap in Fig. 4(a), where the difference between the best accuracy and the worst is 1.68%. As can be seen in Fig. 4(b), we also evaluate the influence of selection rates t. Our method achieves the best performance when $t = 0.7$.

Conclusion

In this paper, combined with the experience of ophthalmologists, a multi-view DR grading framework with coherent diagnostic thought is proposed. This method proposes dynamic selection mechanisms from fine- to coarse-grained features. At the pixel level, we take advantage of the input-dependent selection mechanism of Mamba and the long-distance feature sensitivity of Transformer, to fully learn the features of multi-view fundus images with lesion prompts. At the feature token level, we propose MVTSM to eliminate the negative effects of redundant feature tokens. At the view level, we use MVIMamba to explore the multi-view correlation and use MVMoEM to selectively strengthen important views and weaken conflicting views, to improve the performance of the model. Experimental results show the superiority of our approach over state-of-the-art models.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 82261138629 and 12326610, Guangdong Provincial Key Laboratory under Grant 2023B1212060076, the Foundation for Young Innovative Talents in Ordinary Universities of Guangdong under Grant 2024KQNCX042, the Shenzhen Municipal Science and Technology Innovation Council under Grant JCYJ20240813141424032, JCYJ20240813112420027 and JCYJ20220531101412030, the Stable Support Projects for Shenzhen Higher Education Institutions under grant no. 20231122005530001.

References

Dai, L.; Sheng, B.; Chen, T.; Wu, Q.; Liu, R.; Cai, C.; Wu, L.; Yang, D.; Hamzah, H.; Liu, Y.; et al. 2024. A deep learning system for predicting time to progression of diabetic retinopathy. *Nature Medicine*, 30(2): 584–594.

Davis, J.; and Goadrich, M. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the International Conference on Machine Learning*, 233–240.

Decenciere, E.; Zhang, X.; Cazuguel, G.; Lay, B.; Cochener, B.; Trone, C.; Gain, P.; Ordonez, R.; Massin, P.; Erginay, A.; Charton, B.; and Klein, J.-C. 2014. Feedback on a publicly distributed database: the Messidor database. *Image Analysis Stereology*, 33(3): 231–234.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

EyePACS. 2015. Kaggle-EyePACS. https://www.kaggle.com/c/diabetic-retinopathy-detection/data. Accessed: 2024-08-02.

Fang, X.; Fang, W.; Liu, D.; Qu, X.; Dong, J.; Zhou, P.; Li, R.; Xu, Z.; Chen, L.; Zheng, P.; et al. 2024a. Not all inputs are valid: Towards open-set video moment retrieval using language. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 28–37.

Fang, X.; Liu, D.; Zhou, P.; and Nan, G. 2023. You can ground earlier than see: An effective and efficient pipeline for temporal sentence grounding in compressed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2448–2460.

Fang, X.; Liu, D.; Zhou, P.; Xu, Z.; and Li, R. 2024b. Hierarchical Local-Global Transformer for Temporal Sentence Grounding. *IEEE Transactions on Multimedia*, 26: 3263–3277.

Federation, I. D. 2021. IDF Diabetes Atlas, 10th edn. https://www.diabetesatlas.org. Accessed: 2024-07-27.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Gu, A.; Johnson, I.; Timalsina, A.; Rudra, A.; and Ré, C. 2023. How to train your hippo: State space models with generalized orthogonal basis projections. In *Proceedings of the International Conference on Learning Representations*.

Hafner, D.; Lillicrap, T.; Ba, J.; and Norouzi, M. 2020. Dream to control: Learning behaviors by latent imagination. In *Proceedings of the International Conference on Learning Representations*.

Han, Z.; Zhang, C.; Fu, H.; and Zhou, J. T. 2022. Trusted multi-view classification with dynamic evidential fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(2): 2551–2566.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Hu, J.; Chen, R.; Lu, Y.; Dou, X.; Ye, B.; Cai, Z.; Pu, Z.; and Mou, L. 2019. Single-Field Non-Mydriatic Fundus Photography for Diabetic Retinopathy Screening: A Systematic Review and Meta-Analysis. *Ophthalmic Research*, 62: 61–67.

Huang, Y.; Lyu, J.; Cheng, P.; Tam, R.; and 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.

Jang, E.; Gu, S.; and Poole, B. 2017. Categorical Reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*.

Karafyllis, I.; and Krstic, M. 2011. Nonlinear stabilization under sampled and delayed measurements, and with inputs subject to delay and zero-order hold. *IEEE Transactions on Automatic Control*, 57(5): 1141–1154.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollar, P. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*.

Liu, C.; Wen, J.; Liu, Y.; Huang, C.; Wu, Z.; Luo, X.; and Xu, Y. 2024a. Masked Two-channel Decoupling Framework for Incomplete Multi-view Weak Multi-label Learning. *Advances in Neural Information Processing Systems*, 36.

Liu, C.; Wu, Z.; Wen, J.; Xu, Y.; and Huang, C. 2022a. Localized Sparse Incomplete Multi-view Clustering. *IEEE Transactions on Multimedia*.

Liu, Y.; Gehrig, M.; Messikommer, N.; Cannici, M.; and Scaramuzza, D. 2024b. Revisiting token pruning for object detection and instance segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2658–2668.

Liu, Y.; Zhou, Q.; Wang, J.; Wang, Z.; Wang, F.; Wang, J.; and Zhang, W. 2024c. Dynamic token-pass transformers for semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1827–1836.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Liu, Z.; Mao, H.; Wu, C.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022b. A ConvNet for the 2020s. arXiv:2201.03545.

Luo, X.; Liu, C.; Wong, W.; Wen, J.; Jin, X.; and Xu, Y. 2023. MVCINN: multi-view diabetic retinopathy detection using a deep cross-interaction neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 8993–9001.

Luo, X.; Xu, Q.; Wang, Z.; Huang, C.; Liu, C.; Jin, X.; and Zhang, J. 2024. A Lesion-Fusion Neural Network for Multi-View Diabetic Retinopathy Grading. *IEEE Journal of Biomedical and Health Informatics*, 1(1): 1–10.

Pao, S. I.; Lin, H. Z.; Chien, K. H.; Tai, M. C.; Chen, J. T.; and Lin, G. M. 2020. Detection of Diabetic Retinopathy Using Bichannel Convolutional Neural Network. *Journal of Ophthalmology*, 2020: 1–7.

Riquelme, C.; Puigcerver, J.; Mustafa, B.; Neumann, M.; Jenatton, R.; Susano Pinto, A.; Keysers, D.; and Houlsby, N. 2021. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34: 8583–8595.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Sasaki, Y. 2007. The truth of the F-measure. *Teach Tutor Mater*, 1(5): 1–5.

Srihatrai, P.; and Hlowchitsieng, T. 2018. Thanita The diagnostic accuracy of single- and five-field fundus photography in diabetic retinopathy screening by primary care physicians. *Indian Journal of Ophthalmology*, 66: 94–97.

Sun, R.; Li, Y.; Zhang, T.; Mao, Z.; Wu, F.; and Zhang, Y. 2021. Lesion-Aware Transformers for Diabetic Retinopathy Grading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10938–10947.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the Inception Architecture for Computer Vision. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Trevethan, R. 2017. Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. *Frontiers in Public Health*, 5: 307.

Wilkinson, C. P.; Ferris III, F. L.; Klein, R. E.; Lee, P. P.; Agardh, C. D.; Davis, M.; Dills, D.; Kampik, A.; Pararajasegaram, R.; Verdaguer, J. T.; et al. 2003. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, 110(9): 1677–1682.

Wu, J.; Hu, R.; Xiao, Z.; Chen, J.; and Liu, J. 2021. Vision Transformer-based recognition of diabetic retinopathy grade. *Medical Physics*, 48(12): 7850–7863.

Xie, S.; Girshick, R.; Dollar, P.; Tu, Z.; and He, K. 2017. Aggregated Residual Transformations for Deep Neural Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Xu, Q.; Luo, X.; Huang, C.; Liu, C.; Wen, J.; Wang, J.; and Xu, Y. 2024. HACDR-Net: Heterogeneous-Aware Convolutional Network for Diabetic Retinopathy Multi-Lesion Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6342–6350.

Yuan, Y.; Fu, R.; Huang, L.; Lin, W.; Zhang, C.; Chen, X.; and Wang, J. 2021. Hrformer: High-resolution vision transformer for dense predict. *Advances in neural information processing systems*, 34: 7281–7293.

Zhou, Y.; Chia, M. A.; Wagner, S. K.; Ayhan, M. S.; Williamson, D. J.; Struyven, R. R.; Liu, T.; Xu, M.; Lozano, M. G.; Woodward-Court, P.; et al. 2023. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981): 156–163.

Zhou, Y.; Lei, T.; Liu, H.; Du, N.; Huang, Y.; Zhao, V.; Dai, A. M.; Le, Q. V.; Laudon, J.; et al. 2022. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35: 7103–7114.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. In *Forty-first International Conference on Machine Learning*.