Incomplete Multi-view Diabetic Retinopathy Grading via Self-Supervised Inter- and Intra-View Restoration

Zhihao Wu[1,2], Yuxin Lin[3*], Jie Wen[3*], Wuzhen Shi[4], Linlin Shen[1,2,5]

[1]School of Artificial Intelligence, Shenzhen University
[2]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University
[3]Shenzhen Key Laboratory of Visual Object Detection and Recognition, Harbin Institute of Technology, Shenzhen
[4]College of Electronics and Information Engineering, Shenzhen University
[5]Guangdong Provincial Key Laboratory of Intelligent Information Processing
horatio_ng@163.com, linyuxin6688@gmail.com, jiewen_pr@126.com, wzhshi@szu.edu.cn, llshen@szu.edu.cn

Abstract

Multi-view diabetic retinopathy (DR) grading has achieved remarkable performance by capturing more comprehensive pathological features than single-view methods. However, complete multi-view fundus images are often difficult to obtain in clinical practice, and the performance degrades significantly when fewer views are available. To overcome this limitation, we propose the first incomplete multi-view DR grading framework, aiming to provide accurate diagnosis regardless of the number of available views. It introduces two novel modules. First, cross-view spatial correlation attention (CSCA) captures region correlations across views, automatically identifying and fusing diagnostically relevant spatial features to improve feature representation. Second, self-supervised mask consistency learning (SMCL) formulates a novel pretext task of missing-view information reconstruction by strategically masking inter- and intra-view regions, enabling the model to infer complete features from incomplete views. Benefiting from CSCA and SMCL, our method enhances structural feature consistency across views and effectively compensates for missing information during DR grading. Extensive experiments demonstrate that our method achieves state-of-the-art performance, particularly under realistic conditions where some views are unavailable.

Introduction

Diabetes can precipitate severe complications, profoundly diminishing quality of life (Lee, Wong, and Sabanayagam 2015). Among these, diabetic retinopathy (DR) is a progressive ocular disease that may lead to visual impairment or even irreversible blindness. Therefore, early diagnosis and timely intervention are crucial. However, the demand for DR early screening significantly surpasses the available ophthalmologists' capacity (Mao et al. 2024).

To bridge this gap, automatic DR grading has attracted increasing research attention. It employs deep learning models to capture pathological features of microaneurysms, hemorrhages, and exudates in fundus images, enabling accurate classification of disease severity. Most existing methods rely on single-view fundus images captured at a 45-degree field of view, potentially causing some pathological features to be obscured or missing. To address this problem, multi-view DR grading has emerged as a promising research direction, significantly improving lesion visibility and enabling models to extract more comprehensive cross-view features, thereby achieving more accurate grading (Luo et al. 2023).

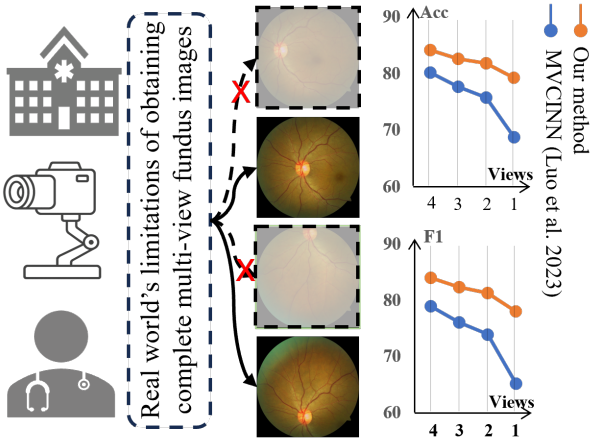

Figure 1: Real-world challenges in acquiring complete multi-view fundus images limit the performance of conventional multi-view DR grading models.

However, multi-view methods are often limited in clinical practice as they rely on a fixed number of fundus views (Wen et al. 2023; Wu et al. 2023; Li et al. 2025, 2023). As shown in Figure 1, in practice, obtaining a complete set of multi-view images is challenging or even infeasible due to patient discomfort, imaging challenges, or equipment limitations. When fewer views than required are available during inference, multi-view model performance declines significantly, sometimes even falling below that of single-view models (see Table 1 and Table 2). In other words, although multi-view DR grading has theoretical advantages, its rigid setting severely undermines its clinical applicability.

To overcome this limitation, we propose the first incom-

*Corresponding authors: Yuxin Lin and Jie Wen.

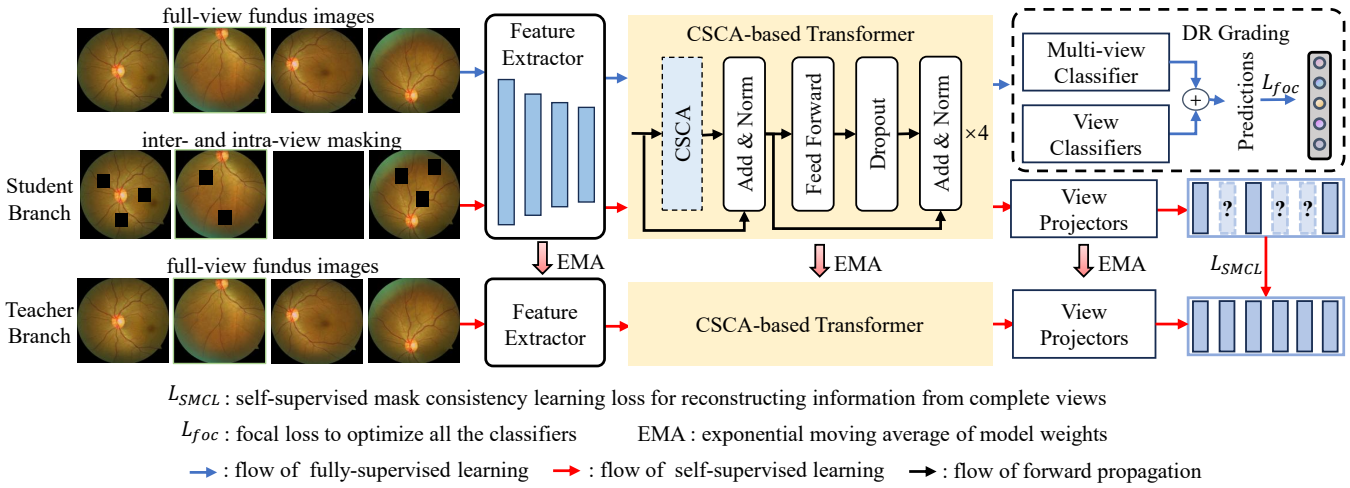

Figure 2: Pipeline of the proposed method.

plete multi-view DR grading framework, which is capable of making accurate grading predictions given a variable number of input views. To this end, we focus on addressing two key challenges faced by complete multi-view approaches: robust feature fusion across variable views as well as effective compensation for inter- and intra-view information loss. We tackle these challenges through two novel modules. First, cross-view spatial correlation attention (CSCA) effectively captures features from each view and enhances cross-view feature fusion by modeling region-wise correlations across different views. Second, self-supervised mask consistency learning (SMCL) introduces a pretext task of missing-view information reconstruction, which leverages a teacher-student model to guide the network in learning complete feature representations from partially missing inter- and intra-view information.

Extensive experimental results demonstrate that our method outperforms all existing single-view and multi-view approaches, especially under missing-view conditions commonly encountered in real-world applications.

To sum up, this work has the following three-fold contributions:

- To the best of our knowledge, this is the first work on incomplete multi-view diabetic retinopathy grading, and it achieves superior performance compared to existing single-view and multi-view models.
- The proposed cross-view spatial correlation attention constructs region-wise correlation matrices and employs dynamic attention to assign high priority to diagnostically consistent features, thereby enabling effective multi-view fusion.
- The proposed self-supervised mask consistency learning employs a teacher-student framework to guide the model in compensating for missing inter- and intra-view information, thereby enabling accurate DR grading under missing-view conditions.

Related Works

Deep learning-based automatic DR grading makes large-scale early screening for diabetes possible. Existing studies can be categorized into single-view and multi-view approaches. Due to factors such as patient cooperation and collection cost, single-view fundus images remain dominant. In the single-view paradigm, researchers focus on developing end-to-end CNN or Transformer architectures to classify DR severity or segment lesion regions. For example, Wang et al. (2022) utilize a pre-trained CNN backbone to extract features for effective severity prediction. Huang et al. (2024) introduce a saliency-guided self-supervised Transformer that improves grading robustness through an auxiliary feature reconstruction task. Xu et al. (2024) propose a heterogeneity-aware CNN that enhances DR lesion segmentation by modeling spatial and channel-wise dependencies. In addition, Huang et al. (2022) design a Transformer model that exploits inter-lesion feature relationships, achieving advanced segmentation performance.

Although single-view DR grading achieves significant breakthroughs, its further development is limited by the inherent constraints of the 45-degree field of view, as pathological features may manifest across broader retinal regions (Lin et al. 2025a). Therefore, multi-view grading emerges as an important research direction, leveraging the complementarity among different views to enhance diagnostic accuracy. Specifically, Luo et al. (2021) pioneer a multi-view deep CNN framework that jointly optimizes feature extraction from both single-view and cross-view perspectives. Building on this, they (Luo et al. 2023) further propose a cross-interaction neural network that integrates local features extracted by CNN with global dependencies modeled by Transformer, effectively capturing lesion details and retinal structural context. In addition, they (Luo et al. 2024) introduce explicit lesion-aware learning, which enhances fine-grained grading performance through an auxiliary lesion segmentation task. These studies highlight the value of multi-view learning in DR grading, but their performance is

heavily constrained by the predefined number of views. Motivated by this, we pioneer the incomplete multi-view DR grading task.

Method

Pipeline

The pipeline of our proposed method is illustrated in Figure 2. In the fully-supervised learning phase, full-view fundus images are fed into a shared backbone to extract features, which are then fused via CSCA-based Transformer and subsequently used for DR grading. In the self-supervised learning phase, the student branch receives randomly masked inputs (simulating incomplete data), while the teacher branch is fed complete-view images, with its parameters updated via EMA (Brotons, Vogels, and Hendrikx 2024). Finally, the self-supervised mask consistency learning loss is applied to enforce consistent feature representations between the student and teacher models.

CSCA-based Transformer

Let $\boldsymbol{\mathcal{V}} = \left[\begin{array}{cccc} \boldsymbol{\mathcal{V}}^1 & \boldsymbol{\mathcal{V}}^2 & \cdots & \boldsymbol{\mathcal{V}}^N \end{array} \right]$ denote the multi-view fundus images, where $\boldsymbol{\mathcal{V}}^i \in \mathbb{R}^{C \times H \times W}$ represents the i-th view of the retina, and N is the number of views. As in previous works (Lu et al. 2023; Wang et al. 2024), the Swin Transformer pretrained on ImageNet (Liu et al. 2021a) is employed as the feature extractor. The extracted multi-view features are denoted as $\boldsymbol{\mathcal{F}} = \left[\begin{array}{cccc} \boldsymbol{\mathcal{F}}^1 & \boldsymbol{\mathcal{F}}^2 & \cdots & \boldsymbol{\mathcal{F}}^N \end{array} \right]$.

Similar to Vision Transformer (ViT) (Dosovitskiy et al. 2020), the feature maps $\boldsymbol{\mathcal{F}}^i$ are reshaped into sequential representations. Let $ToLN\left(\bullet\right)$ denote the operation that partitions the feature maps and performs linear projection. The resulting token set can be expressed as:

$$\mathbf{T}^i = ToLN\left(\boldsymbol{\mathcal{F}}^i\right) = \left[\begin{array}{cccc} \mathbf{T}_1^i & \mathbf{T}_2^i & \cdots & \mathbf{T}_L^i \end{array} \right]^\top + E_{pos}, \tag{1}$$

where L and E_{pos} represent the number of patches and the position embedding, respectively. Notably, $\mathbf{T}_j^i$ corresponds to the j-th token, which encapsulates the j-th region of the i-th view. Since conventional multi-view DR grading methods rely on pooling (Luo et al. 2021), they typically fail to preserve spatial correlations between 2D representations of the same 3D retina, leading to misalignment and ineffective information integration across views.

To break this limitation, we propose CSCA, which explicitly encodes the relevance between regions across different views into the learnable region-wise correlation matrices $\mathbf{M}^{ij}$. Specifically, the token sequence $\mathbf{T}^i$ is projected into a normalized query matrix $\mathbf{Q}^i$, key matrix $\mathbf{K}^i$ and value matrix $\mathbf{V}^i$. Then the correlation matrix between the i-th and j-th view can be computed as:

$$\mathbf{M}^{ij} = \mathbf{Q}^i \left(\mathbf{K}^j\right)^\top. \tag{2}$$

It is worth noting that each row of $\mathbf{Q}^i$ and $\mathbf{K}^j$ represents a normalized feature vector corresponding to a specific region in their respective views. This matrix multiplication effectively computes cosine similarity coefficients between regions across views, quantifying their spatial and semantic relevance. Formally, each element (m, n) in the learnable cross-view correlation matrix $\mathbf{M}^{ij} \in \mathbb{R}^{L \times L}$ quantifies the dynamic relevance between the m-th region in the i-th view and the n-th region in the j-th view. This facilitates precise alignment of spatial and semantic information across multiple views.

To bridge the spatial and contextual disparities inherent in multi-perspective representations, CSCA is computed as follows:

$$\mathbf{CSCA}^i = \frac{\sum_{j=1}^{N} FC\left(Softmax\left(\frac{\mathbf{M}^{ij}}{\sqrt{D}}\right)\mathbf{V}_i\right)}{N}, \tag{3}$$

where FC denotes a fully connected layer applied to the weighted views, and D is the token embedding dimension. It is crucial for enhancing feature fusion across varying perspectives by effectively integrating information from different views while prioritizing the most diagnostically relevant regions. By dynamically adjusting the contribution of each view based on its relevance, the CSCA module enables more accurate and robust multi-view learning. Even when some views are incomplete or missing, the model can rely on the available views to extract comprehensive and discriminative features, which are essential for precise DR grading.

Finally, the CSCA-based Transformer is composed of 4 CSCA modules, each followed by Add&Norm, a feed-forward layer, dropout, and another Add&Norm layer: $\tilde{\boldsymbol{\mathcal{F}}}^i = Tr_{CSCA}(\boldsymbol{\mathcal{F}}^i)$, where $\tilde{\boldsymbol{\mathcal{F}}}^i$ represents the output features. In this work, we employ the multi-head version of the forward-attention module.

SMCL

SMCL consists of a student branch and a teacher branch, both sharing the same architecture, which includes a feature extractor, a CSCA-based Transformer, and a view projector. The student branch is trained to extract detailed and discriminative features for incomplete multi-view fundus images, leveraging both intra-view and inter-view consistency learning. Intra-view consistency is achieved by randomly masking regions, while inter-view consistency is enforced by training on masked views. This dual consistency learning encourages the student branch to reconstruct missing features based on the teacher's output from complete multi-view images. Therefore, our method can generalize across varying numbers of views.

Masking To perform intra-masking within each view, we first partition all views into regular, non-overlapping patches. A subset of these patches is randomly selected and retained, while the remaining patches are masked, resulting in a partially masked multi-view fundus image set. In the inter-view masking process, we randomly select between 0 and $(N-1)$ views to be entirely masked. These processes mimic clinical settings where fundus image counts per patient vary. The masked multi-view fundus images are denoted as $\hat{\boldsymbol{\mathcal{V}}} = \left[\begin{array}{cccc} \hat{\boldsymbol{\mathcal{V}}}^1 & \hat{\boldsymbol{\mathcal{V}}}^2 & \cdots & \hat{\boldsymbol{\mathcal{V}}}^N \end{array} \right]$.

SMCL Loss We define the SMCL loss as the expected distance between the features from the student $Stu(*)$ and

teacher $Tea(*)$ branches:

$$L_{SMCL}(\theta) = \sum_{i=1}^{N} \left\| Stu(\hat{\boldsymbol{\mathcal{V}}}^i, \theta_{stu}) - Tea(\boldsymbol{\mathcal{V}}^i, \theta_{tea}) \right\|^2 = \sum_{i=1}^{N} \left\| Pro_{stu}\left(\tilde{\boldsymbol{\mathcal{F}}}^i_{stu}\right) - Pro_{tea}\left(\tilde{\boldsymbol{\mathcal{F}}}^i_{tea}\right) \right\|^2, \tag{4}$$

where θ_{stu} and θ_{tea} denote the network weights of the student and teacher branches, respectively. Pro_{stu} and Pro_{tea} represent the view projectors corresponding to each branch. θ_{tea} is updated using exponential moving average (EMA) (Brotons, Vogels, and Hendrikx 2024):

$$\theta^t_{tea} = \alpha\theta^{t-1}_{tea} + (1-\alpha)\theta^t_{stu}, \tag{5}$$

where θ^t denotes the weights at training step t, and α is a smoothing coefficient hyperparameter. This approach ensures that the teacher's outputs at each step are formed as an ensemble of its current and historical states, effectively capturing long-term knowledge. By leveraging this temporal ensembling, the teacher branch progressively aggregates grading-guided information, enhancing its ability to generate more accurate and diagnostically relevant feature representations.

DR Grading

As shown in Figure 2, DR grading predictions are achieved by integrating a multi-view classifier (MVC) and N view classifiers (VCs). MVC operates on fused features from all N views using CSCA-based Transformer, while each VC performs classification independently based on features extracted from a single view. Let the classification token corresponding to the i-th view from CSCA-based Transformer be denoted as cls^i. The fused prediction from MVC is then computed as:

$$P^{mv} = MVC\left(stack\left(Tr\left(\left[\begin{array}{cccc} cls^1 & cls^2 & \cdots & cls^N \end{array}\right]\right)\right)\right) \in [0,1]^C, \tag{6}$$

where Tr refers to Transformer encoder in ViT (Dosovitskiy et al. 2020). Specifically, CSCA produces one classification token (global summary) per view, and Tr takes these tokens as input to model global-level relationships across views. Similarly, the classification result for the i-th view using the individual VC is computed as:

$$P^i = VC^i\left(cls^i\right) \in [0,1]^C. \tag{7}$$

Training In practice, high-grade DR samples are rare, leading to a class imbalance issue. To address this, we introduce focal loss (Lin et al. 2017) during training:

$$L_{foc} = -\sum_{j\in\{1,\ldots,N,mv\}} \left(1-P^j_k\right)^{\gamma} \log\left(P^j_k\right), \tag{8}$$

where γ is a tunable parameter that down-weights easy samples, indirectly giving rarer classes a higher relative impact during training. p_k denotes the predicted probability for the k-th grade.

The final loss is the sum of L_{SMCL} and L_{foc}:

$$L = L_{SMCL} + L_{foc}. \tag{9}$$

Inference Due to the potential absence or sparse diagnostic information in certain views during inference, assigning adaptive weights to each classifier is crucial for effective fusion. Let A and M denote the index sets of available input views and missing views, respectively. The final prediction score is computed as:

$$Ps = \frac{w^{mv}P^{mv} + \sum_{i\in A} w^i P^i + \sum_{k\in M} w^k P^k}{N+1}. \tag{10}$$

To ensure that Ps remains within the probability range $[0,1]^C$, the sum of the assigned weights must satisfy the following constraint:

$$w^{mv} + \sum_{i\in A} w^i + \sum_{k\in M} w^k = N+1. \tag{11}$$

For complete multi-view DR grading, all N views are available, and each view is equally important. Since no views are missing, the missing view set M is empty ($M = \emptyset$), and the final grading result is computed using uniform weighting:

$$w^{mv} = w^i = 1,\ \forall i \in A. \tag{12}$$

For incomplete multi-view DR grading, the number of input views ranges from 1 to N, with some views missing. In this case, assigning higher weights to available views and lower weights to missing ones is a highly intuitive strategy. To this end, we define the weights for the MVC and the classifiers corresponding to the missing views as follows:

$$w^{mv} = w^k = \frac{1}{2},\ \forall k \in M. \tag{13}$$

For the classifiers corresponding to available views, the weights are computed as:

$$w^i = \frac{N+1-w^{mv}-\sum_{k\in M} w^k}{N_A},\ \forall i \in A, \tag{14}$$

where N_A represent the number of available views. This adaptive weighting mechanism ensures that available views contribute more significantly to grading.

Experiments

Datasets and Evaluation Metrics

We conduct experiments on the four-view DR grading dataset (Luo et al. 2021), which consists of 25,848 training images and 8,604 testing images. Following previous studies (Luo et al. 2023; Lin et al. 2025a), we evaluate model performance using four standard classification metrics, including precision (Pre), F1-score (F1), specificity (Spec), and accuracy (Acc).

Implementation Details

For CSCA-based Transformer and each view projector, the depth and the number of heads are set to 4 and 8, respectively. For intra-view masking, fundus images are divided into non-overlapping patches of size 16×16, and 25% of

Models	Pre	F1	Spec	Acc
VGG19	71.3	70.7	73.6	74.1
ResNet50	68.1	68.5	71.3	71.9
ResNet101	69.8	69.4	71.6	73.1
ResNext50	67.8	68.0	70.1	72.0
Swin-T	67.7	67.8	70.3	71.7
Swin-S	69.3	69.4	71.9	73.0
Swin-B	71.1	70.4	72.4	74.1
ConvNeXt-T	70.3	69.8	72.7	73.4
ConvNeXt-S	70.6	69.9	72.6	73.6
ConvNeXt-B	72.0	71.1	73.9	74.4
DCDR	64.6	63.4	65.9	69.8
HIDIF	73.2	72.9	76.4	75.4
Ours	**77.8**	**78.0**	**82.5**	**79.2**

Table 1: Quantitative comparison with single-view methods.

the patches are randomly masked. For inter-view masking, the number of masked views is determined using random.randint, and the masked view indices are selected using random.shuffle. Both operations follow uniform distributions.

During training, RandomResizedCrop, RandomHorizontalFlip, and RandomVerticalFlip are applied for data augmentation. We use the Adam optimizer with an initial learning rate of $5e^{-6}$, and train the model for 100 epochs with a batch size of 4. The smoothing coefficient used in EMA is set to 0.999. For the focal loss, γ is set to 2 (Lin et al. 2017).

Comparison with Single-view Models

To demonstrate the effectiveness of our approach with single-view inputs during inference, we compare it with commonly-used classification models, including VGG19 (Simonyan 2014), ResNet50 (He et al. 2016), ResNet101(He et al. 2016), ResNext50 (Xie et al. 2017), Swin-T (Liu et al. 2021b), Swin-S (Liu et al. 2021b), Swin-B (Liu et al. 2021b), ConvNeXt-T (Liu et al. 2022), ConvNeXt-S (Liu et al. 2022), and ConvNeXt-B (Liu et al. 2022), as well as single-view DR grading methods DCDR (Bosale et al. 2024) and HIDIF (Lin et al. 2025b). For a fair comparison, all these models are trained on full-view data, treating each four-view sample as four single-view instances, as described in (Luo et al. 2023).

As shown in Table 1, our method outperforms all single-view models across all metrics. This superior performance stems from the fact that single-view methods treat multi-view data as independent images, neglecting the inherent relationships between views during training. In contrast, our method can infer missing-view information by leveraging the proposed self-supervised mask consistency learning.

Comparison with Multi-view Models

We further compare our method with state-of-the-art multi-view DR grading models (see Table 2), including MVCINN (Luo et al. 2023), LFNN (Luo et al. 2024), CVSA (Lin et al. 2025a), SMVDR (Luo et al. 2025), MVCNN (Su et al. 2015), and WGLIM (Hu et al. 2025). Given the common occurrence of missing views in real-world applications, we

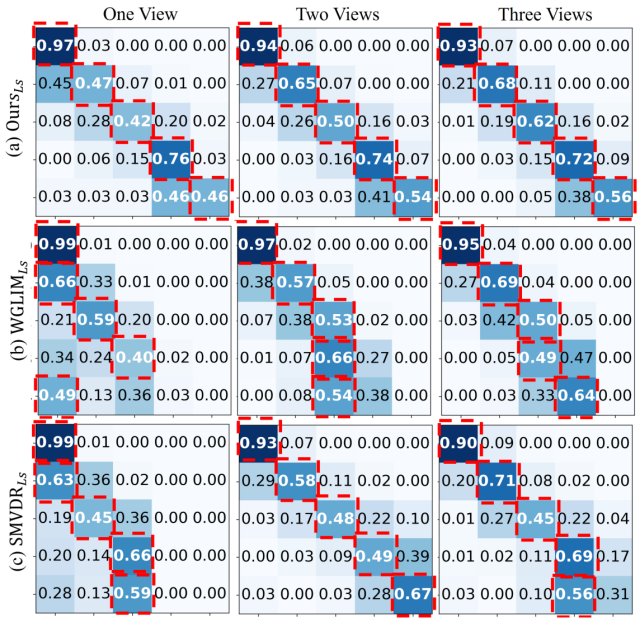

Figure 3: Confusion matrices of the comparison methods under incomplete-view scenarios.

conduct comparative evaluations under both full-view and partial-view conditions. As the performance is minimally affected by the specific choice of available views, we report the results using the sequential selection of one to three views. It can be observed that our models outperform competing methods on most metrics. Notably, even without incorporating lesion information, our method achieves the highest or near-highest performance, outperforming all previous methods. This result is surprising and significant, as it demonstrates that our method can learn highly discriminative features and effectively compensate for missing lesion information through the multi-view reasoning and self-supervised consistency.

We further show the confusion matrices of two state-of-the-art models, SMVDR$_{Ls}$ and WGLIM$_{Ls}$, compared to our lesion-assisted model (Ours$_{Ls}$) under incomplete multi-view conditions (see Figure 3). In each matrix, the red dashed box highlights the highest value in each row, corresponding to the grade that receives the majority of predictions. This visualization provides insights into the model's tendency to misclassify and its ability to handle samples from each DR grade. Across all three scenarios, our model consistently places the red boxes along the main diagonal, indicating that it correctly predicts the majority of samples for each grade. In contrast, WGLIM$_{Ls}$ fails to correctly classify any samples from Grade 4 across all scenarios, misclassifying them into lower grades. Similarly, SMVDR$_{Ls}$ exhibits severe misclassification in the one-view scenario, where samples from both Grade 3 and Grade 4 are entirely misclassified. Even in the three-view scenario, SMVDR$_{Ls}$ struggles to distinguish Grade 4 from Grade 3, indicating persistent confusion between adjacent severity levels. These misclassifications are largely attributed to the inherent

Models	One View				Two Views				Three Views				Four Views			
	Pre	F1	Spec	Acc	Pre	F1	Spec	Acc	Pre	F1	Spec	Acc	Pre	F1	Spec	Acc
MVCINN	65.3	65.1	68.8	68.7	73.5	73.9	78.3	75.7	76.0	76.0	80.6	77.6	78.9	78.9	83.3	80.1
LFNN$_{Ls}$	68.5	51.9	**90.4**	72.9	71.0	59.0	86.0	75.8	78.1	77.6	81.9	79.3	81.5	81.3	86.9	82.2
CVSA	75.8	46.8	90.2	40.5	79.3	76.8	84.7	78.2	79.9	79.5	85.1	80.6	82.1	81.9	81.9	82.6
SMVDR$_{Ls}$	63.4	66.4	70.9	71.8	78.9	78.4	87.7	78.3	80.5	80.1	90.2	80.0	82.2	83.7	91.3	84.0
MVCNN_V	70.6	70.6	77.2	71.8	74.8	75.1	80.0	76.8	75.4	75.3	78.7	77.6	77.6	77.2	80.5	79.1
MVCNN_R	68.8	69.3	72.4	72.3	72.2	72.2	75.3	75.1	73.7	73.3	75.8	76.4	75.8	75.1	79.2	77.4
WGLIM$_{Ls}$	76.7	76.6	84.0	78.4	79.1	79.7	88.0	80.8	79.1	79.8	88.0	80.8	83.9	83.6	90.0	84.1
Ours	**77.8**	**78.0**	82.5	79.2	81.0	81.3	87.2	81.8	82.3	82.3	89.4	82.5	84.0	84.0	90.2	84.1
Ours$_{Ls}$	**77.8**	77.8	81.7	**79.5**	**81.6**	**81.9**	**88.3**	**82.2**	**83.0**	**83.0**	**90.6**	**82.9**	**84.8**	**84.7**	**91.8**	**84.7**

Table 2: Quantitative comparison with multi-view methods. "_V" and "_R" denote using VGG19 and ResNet50 as backbones, respectively. The subscript $_{Ls}$ indicates the use of lesion segmentation masks as additional input, achieved through Retinal Vessel Reinforce (Lin et al. 2025a).

Models	Grade 0			Grade 1			Grade 2			Grade 3			Grade 4		
	F1	Pre	Spec	F1	Pre	Spec	F1	Pre	Spec	F1	Pre	Spec	F1	Pre	Spec
MVCINN	91.3	86.7	75.9	56.4	68.3	94.1	59.3	57.4	95.8	68.1	70.0	97.9	44.8	68.4	99.7
LFNN$_{Ls}$	92.4	89.7	82.1	66.3	69.5	92.7	59.0	62.1	96.8	70.9	69.5	97.6	17.0	50.0	99.9
CVSA	92.3	89.2	81.2	62.6	73.6	**95.0**	64.2	61.0	96.0	73.2	72.7	98.0	**64.1**	64.1	99.3
SMVDR$_{Ls}$	93.5	93.5	90.2	71.7	71.2	90.0	60.3	60.0	95.1	74.2	69.4	97.9	30.43	**99.9**	**100.0**
MVCNN_V	90.1	84.5	71.1	50.0	65.3	94.3	60.2	**65.3**	**97.3**	73.6	66.8	97.0	38.5	76.9	99.8
MVCNN_R	89.4	83.6	69.3	46.1	62.1	94.1	59.4	58.1	95.9	68.4	66.0	97.3	22.2	83.3	99.9
WGLIM$_{Ls}$	93.5	92.3	87.0	71.4	71.0	92.3	59.9	63.9	97.1	**74.7**	71.9	97.7	29.8	87.5	99.9
Ours	93.3	92.4	87.4	70.2	71.2	92.7	62.4	60.9	96.2	74.0	**75.0**	**98.2**	62.3	86.4	99.9
Ours$_{Ls}$	**93.9**	**93.7**	**89.7**	**72.3**	**73.7**	93.4	**65.1**	62.2	96.1	71.7	73.2	98.1	55.7	61.1	99.2

Table 3: Grade-wise performance analysis under the complete-view setting.

class imbalance in real-world DR datasets, where high-grade cases (Grades 3 and 4) are significantly underrepresented compared to lower grades. In contrast, the superior diagonal alignment in confusion matrices of our model demonstrates its effectiveness in mitigating the impact of data imbalance.

In addition, Table 3 presents a comprehensive performance comparison across DR grades under the complete-view setting. For Grade 0 (healthy), both Ours and Ours$_{Ls}$ achieve outstanding results, with Ours$_{Ls}$ attaining the highest specificity, F1 score, and precision, indicating strong reliability in identifying non-DR cases. For Grades 1 to 4, our models maintain superior or highly competitive results. Although SMVDR$_{Ls}$ reports the highest specificity and precision for Grade 4, its F1 score remains notably low (30.43), reflecting poor recall and failure to adequately identify true positives. This imbalance illustrates a critical weakness in handling underrepresented classes. In contrast, our model maintains high F1 scores across all grades even in the absence of lesion information, demonstrating its robustness, discriminative power, and clinical applicability.

Ablation Study

We also conduct a series of ablation experiments to evaluate the contribution of SMCL and CSCA (Figure 4) and analyze the impact of masking ratio (Figure 5). All models are tested under different numbers of input views to thoroughly assess the robustness and effectiveness.

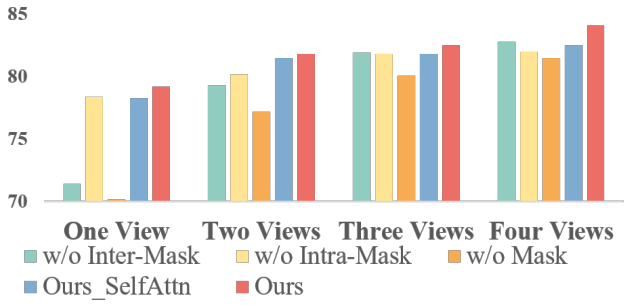

Figure 4: Ablation study evaluating the impact of each proposed module on DR grading accuracy across varying numbers of input views.

Impact of SMCL As shown in Figure 4, to evaluate the effectiveness of SMCL, we design three ablation settings: (1) training with only self-supervised inter-view mask consistency learning ("w/o Intra-Mask"), (2) training with only self-supervised intra-view mask consistency learning ("w/o Inter-Mask"), and (3) training without both masking strategies ("w/o Mask"). The complete version is referred to as "Ours".

We can see that (1) Models without inter-view masking and without both masking demonstrate significantly lower grading accuracies, especially when only one view is avail-

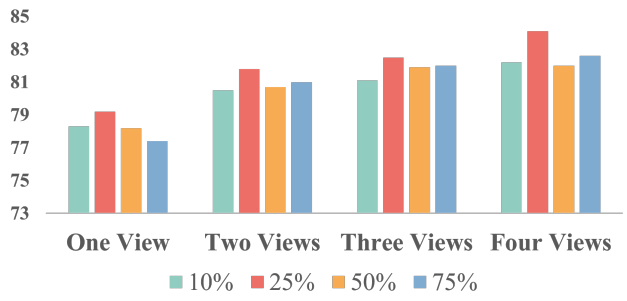

Figure 5: Ablation study evaluating the impact of masking ratio.

Models	Pre	F1	Spec	Acc
DINO	82.6	81.3	**90.5**	81.7
SimMIM	82.8	82.6	89.4	82.9
SMCL	**84.0**	**84.0**	90.2	**84.1**

Table 4: Quantitative comparison between SMCL and other self-supervised learning methods.

able. This highlights the importance of self-supervised inter-view mask consistency learning, which allows the model to infer and reconstruct missing-view features from available inputs, maintaining higher performance under incomplete-view conditions. (2) Compared to "w/o Intra-Mask", the full model achieves higher accuracy. This indicates that self-supervised intra-view mask learning can enhance feature representation.

Additionally, we experiment with Swin-B pre-trained with DINO (Caron et al. 2021) and SimMIM (Xie et al. 2022), without using SMCL. As shown in Table 4, while both methods perform well under the complete-view condition, our SMCL slightly outperforms them, further validating the advantage of learning both inter- and intra-view consistency.

Impact of CSCA To assess the impact of CSCA, we replace it with a conventional ViT module, maintaining the same number of heads and feature dimensions for a fair comparison. This variant is referred to as "Ours_SelfAttn" in Figure 4. We can see that it leads to a significant performance drop, highlighting the critical role of CSCA in effectively modeling inter-view region-wise correlations. These correlations are essential for integrating complementary information from different views, thereby enabling more accurate and robust DR grading.

Impact of Masking Ratio As shown in Figure 5, a masking ratio of 25% consistently achieves the highest grading accuracy across all input view settings. This suggests that a moderate masking level strikes an optimal balance between information retention and self-supervised learning. Too low a masking ratio may hinder the model's ability to learn robust feature reconstruction, while too high a ratio may remove excessive information, impeding effective learning.

Models	FLOPs	Pre	F1	Spec	Acc
ConvNeXt-Tiny	4.46G	83.2	82.4	89.6	83.0
PVT-V2-B0	0.53G	83.4	82.6	90.1	83.1
Swin-B	10.22G	**84.0**	**84.0**	**90.2**	**84.1**

Table 5: Quantitative results with different backbones.

Practical Applicability

To evaluate the applicability of the proposed method in a hospital setting, we test inference time of the model with Swin-B backbone. Specifically, the inference time per four-view sample is 0.00593s on an NVIDIA 3090 GPU or 0.261s on an Intel i9-12900 CPU. This result suggests that our method can be deployed effectively in hospital environments, where quick and reliable predictions are crucial.

We also explore the use of more lightweight backbones to reduce computational overhead. As shown in Table 5, lightweight backbones result in slight performance drops under full-view inference. This demonstrates that our method is not sensitive to the backbone. Thus, the ability to maintain high performance with lightweight backbones further emphasizes the practical applicability of our method in real-world medical environments.

In addition, to further validate the applicability of our method, we evaluate it on the two-view DR grading dataset DRTiD (Hou et al. 2022). Under fair comparison, our method achieves 7.54% higher accuracy than the original approach. The superiority on the multi-source multi-view DR grading datasets further demonstrates the reliability of the proposed method in real-world applications.

Conclusion

In this paper, we propose the first incomplete multi-view DR grading model, specifically designed to overcome the rigidity of previous multi-view models, which require a fixed number of input views. It dynamically adapts to an arbitrary number of fundus views, significantly enhancing clinical practicality by effectively handling scenarios with incomplete or limited imaging data. Benefiting from the proposed CSCA and SMCL, our framework effectively captures critical pathological features across views and compensates for missing visual information. Extensive experiments demonstrate that our model surpasses both single-view approaches and state-of-the-art multi-view methods, particularly excelling under challenging conditions with fewer available views. Remarkably, even in scenarios with only a single available view, our model outperforms existing single-view methods, showcasing its robustness. Overall, this study significantly advances automated DR diagnosis, highlighting a path toward more flexible, accurate, and clinically viable deep learning applications in ophthalmology. In future work, we will explore the potential of the proposed model in a broader range of retinal imaging tasks and further validate the advantage of multi-view knowledge transfer in single-view clinical scenarios.

Acknowledgments

This work was supported in part by the Scientific Foundation for Youth Scholars of Shenzhen University under Grant 868-000001033387, the Guangdong Basic and Applied Basic Research Foundation under Grant No. 2024A1515030213, the National Natural Science Foundation of China under Grant 62372136, 82261138629 and 12326610, the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University under Grant No.VRLAB2025B01, the Shenzhen Science and Technology Program under Grant JCYJ20240813141358076, and in part by the Guangdong Provincial Key Laboratory under Grant 2023B1212060076.

References

Bosale, A. A.; et al. 2024. Detection and classification of diabetic retinopathy using deep learning algorithms for segmentation to facilitate referral recommendation for test and treatment prediction. *arXiv preprint arXiv:2401.02759*.

Brotons, D. M.; Vogels, T.; and Hendrikx, H. 2024. Exponential moving average of weights in deep learning: Dynamics and benefits. *Transactions on Machine Learning Research Journal*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hou, J.; Xu, J.; Xiao, F.; Zhao, R.-W.; Zhang, Y.; Zou, H.; Lu, L.; Xue, W.; and Feng, R. 2022. Cross-field transformer for diabetic retinopathy grading on two-field fundus images. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 985–990. IEEE.

Hu, Y.; Lin, Y.; Liu, C.; Luo, X.; Dou, X.; Xu, Q.; and Xu, Y. 2025. Wavelet-based Global-Local Interaction Network with Cross-Attention for Multi-View Diabetic Retinopathy Detection. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 1–6.

Huang, S.; Li, J.; Xiao, Y.; Shen, N.; and Xu, T. 2022. RTNet: relation transformer network for diabetic retinopathy multi-lesion segmentation. *IEEE Transactions on Medical Imaging*, 41(6): 1596–1607.

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.

Lee, R.; Wong, T. Y.; and Sabanayagam, C. 2015. Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. *Eye and vision*, 2: 1–25.

Li, X.; Pan, Y. P.; Sun, Y.; Sun, Q.; Sun, Y.; W. Tsang, I.; and Ren, Z. 2025. Incomplete Multi-view Clustering with Paired and Balanced Dynamic Anchor Learning. *IEEE Transactions on Multimedia*, 7087–7098.

Li, X.; Sun, Y.; Sun, Q.; Ren, Z.; and Sun, Y. 2023. Cross-view graph matching guided anchor alignment for incomplete multi-view clustering. *Information Fusion*, 100: 101941.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Lin, Y.; Dou, X.; Luo, X.; Wu, Z.; Liu, C.; Luo, T.; Wen, J.; kuen Ling, B. W.; Xu, Y.; and Wang, W. 2025a. Multi-view diabetic retinopathy grading via cross-view spatial alignment and adaptive vessel reinforcing. *Pattern Recognition*, 111487.

Lin, Y.; Wang, W.; Luo, X.; Wu, Z.; Liu, C.; Wen, J.; and Xu, Y. 2025b. Deep Hierarchies and Invariant Disease-Indicative Feature Learning for Computer Aided Diagnosis of Multiple Fundus Diseases. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5325–5333.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021a. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. 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.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986.

Lu, L.; Pan, X.; Jin, P.; and Ding, Y. 2023. Swin-mmc: Swin-based model for myopic maculopathy classification in fundus images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 18–30. Springer.

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.; Pu, Z.; Xu, Y.; Wong, W. K.; Su, J.; Dou, X.; Ye, B.; Hu, J.; and Mou, L. 2021. MVDRNet: Multi-view diabetic retinopathy detection by combining DCNNs and attention mechanisms. *Pattern Recognition*, 120: 108104.

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*, 1–11.

Luo, X.; Xu, Q.; Wu, H.; Liu, C.; Lai, Z.; and Shen, L. 2025. Like an Ophthalmologist: Dynamic Selection Driven Multi-View Learning for Diabetic Retinopathy Grading. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1–9.

Mao, J.; Ma, X.; Bi, Y.; and Zhang, R. 2024. A Comprehensive Federated Learning Framework for Diabetic Retinopathy Grading and Lesion Segmentation. *IEEE Transactions on Big Data*.

Simonyan, K. 2014. Very deep convolutional networks for large-scale image recognition.

Su, H.; Maji, S.; Kalogerakis, E.; and Learned-Miller, E. 2015. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, 945–953.

Wang, J.; Mao, Y.-A.; Ma, X.; Guo, S.; Shao, Y.; Lv, X.; Han, W.; Christopher, M.; Zangwill, L. M.; Bi, Y.; et al. 2024. ODFormer: Semantic fundus image segmentation using transformer for optic nerve head detection. *Information Fusion*, 112: 102533.

Wang, X.; Xu, M.; Zhang, J.; Jiang, L.; Li, L.; He, M.; Wang, N.; Liu, H.; and Wang, Z. 2022. Joint Learning of Multi-Level Tasks for Diabetic Retinopathy Grading on Low-Resolution Fundus Images. *IEEE Journal of Biomedical and Health Informatics*, 26(5): 2216–2227.

Wen, J.; Liu, C.; Deng, S.; Liu, Y.; Fei, L.; Yan, K.; and Xu, Y. 2023. Deep double incomplete multi-view multi-label learning with incomplete labels and missing views. *IEEE transactions on neural networks and learning systems*.

Wu, L.; Zhang, Q.; Hou, J.; and Xu, Y. 2023. Leveraging single-view images for unsupervised 3D point cloud completion. *IEEE Transactions on Multimedia*.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1492–1500.

Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; and Hu, H. 2022. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9653–9663.

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.