SIGraph: Saliency Image-Graph Network for Retinal Disease Classification in Fundus Image

Peng Zhang[1*], Yuan Li[1*], Haotian Song[1], Yankai Jiang[1], Yubo Tao[1†], Hai Lin[1†], Hongguang Cui[2]

[1]State Key Laboratory of CAD&CG, Zhejiang University, China
[2]The First Affiliated Hospital, Zhejiang University School of Medicine, China
zhpeng980@zju.edu.cn, yuanli@zju.edu.cn, hasong@zju.edu.cn, jyk1996ver@zju.edu.cn, taoyubo@cad.zju.edu.cn, lin@cad.zju.edu.cn, 1189002@zju.edu.cn

Abstract

An efficient and precise diagnosis of retinal diseases is a fundamental goal for auxiliary diagnostic systems in ophthalmology. Inspired by the importance of scattered subtle lesions in manual retinal disease diagnosis, recent research has achieved state-of-the-art performance by mining information related to subtle lesions, including their texture and shape. However, the spatial distribution patterns of subtle lesion areas, which are also crucial in manual diagnosis, have been overlooked in existing research. Neglecting these spatial distribution patterns (e.g., the ring distribution of microaneurysms in diabetic macular edema) may negatively impact the diagnostic process. In this paper, we introduce the **S**aliency **I**mage-**Graph** (SIGraph) network to capture the spatial distribution patterns of lesion areas. We first employ saliency-based perception to identify latent lesion pixels. Subsequently, we propose a novel image-graph block to efficiently capture the global distribution of abundant lesion pixels with minimal information loss. By leveraging additional distribution patterns, SIGraph achieves state-of-the-art performance with at least a 1.5% performance gain across three datasets. Furthermore, ablation studies demonstrate that our image-graph block can be integrated into other visual backbones and effectively boost performance.

Introduction

The prevalence of visual disorders such as glaucoma, diabetic retinopathy, and age-related macular degeneration is increasing as the global population ages. In this context, numerous deep-learning-based methods (Li et al. 2019; Wang et al. 2017; Han et al. 2021; Li et al. 2019; Wang et al. 2017; Han et al. 2021) have been developed to achieve high-precision automatic diagnosis. However, recent research has demonstrated that directly employing networks designed for general visual tasks may yield comparatively inferior results when applied to retinal disease classification. These networks often overlook small, scattered lesions distributed across the entire retina, such as microaneurysms and hemorrhages. In clinical practice, ophthalmologists routinely identify retinal diseases in fundus images by carefully examining these subtle features. Certain studies have developed saliency-based models (Jiang et al. 2022; Demidov et al. 2023) to extract features from subtle lesion areas. The key to these methods lies in utilizing saliency maps to match the distribution of lesions, directing the network's focus towards these subtle areas and enhancing diagnostic accuracy. Although these methods have demonstrated effectiveness, they overlook a critical aspect of retinal diseases: **the spatial distribution pattern of lesion areas**, i.e., the relative topological relationships of these lesions. Studies (Ong et al. 2023; Walter et al. 2002) have shown that specific retinal diseases tend to appear in particular areas of the retina, displaying distinct patterns of lesion distribution. For example, in diabetic macular edema, hard exudates often appear in isolated stripes or clusters or around large, ring-distributed microaneurysms. Therefore, a comprehensive understanding of the specific distribution patterns of lesion areas associated with retinal diseases is essential for accurate diagnoses.

Graph-based methods have demonstrated efficacy in modeling topological data relationships. Recent advancements have introduced Graph Neural Networks (GNNs) to ophthalmology. For instance, DRG-NET (Feng et al. 2023) utilized GNN for diabetic retinopathy grading. Other approaches (Feng et al. 2023; Hu et al. 2023) integrate GNNs with Convolutional Neural Networks (CNNs) for retinopathy classification. However, these methods have not fully exploited the correlation between lesion distribution patterns and associated diseases. Moreover, they often neglect the interconnections among subtle lesion areas in retinopathy, potentially compromising diagnostic accuracy.

Building on these insights, we propose **SIGraph**, a novel approach to incorporate the spatial distribution patterns of subtle scattered lesion areas into the diagnostic process. Our method enhances the granularity of lesion areas from image **patches** to **pixels** and constructs a lesion graph on these pixels to capture the spatial distribution patterns of lesion areas. However, constructing a complete graph on all latent lesion pixels to obtain global distribution information incurs substantial computational costs. Utilizing neighborhood-based graph construction algorithms, such as k-nearest neighbors (k-NNs), may lead to graph disconnection, thereby losing the global perception of lesion distribution. To address this challenge, we propose an efficient graph construction algorithm based on the **minimum Manhattan distance in four quadrants**, which ensures global connectivity while mini-

[*]These authors contributed equally.
[†]Corresponding author.

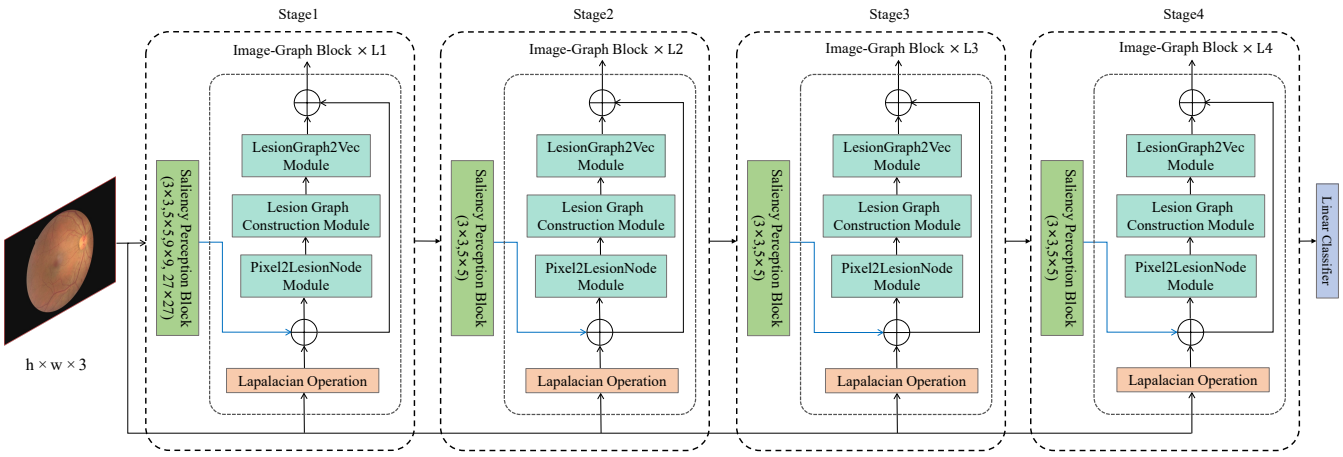

Figure 1: The architecture of SIGraph. It comprises four consecutive stages, each including a visual block and a sequence of image-graph blocks.

mizing edge connections. Finally, to assess the global distribution patterns of lesion areas on the lesion graph, commonly used approaches involve message-passing and pooling operations. However, the effectiveness of direct use of these approaches is limited when applied to our constructed lesion graph, which is characterized by a large number of nodes and sparse edges. To address this challenge, we propose a **mixed graph pooling operation** to capture the distribution pattern of lesion areas while maintaining computational efficiency. Our main contributions are summarized as follows:

- We propose a novel framework for diagnosing retinal diseases, which effectively integrates the spatial distribution of subtle scattered lesion areas into the diagnosis process.
- We propose a new image-graph block to effectively capture the global distribution patterns of numerous lesion pixels. This block is not only applicable to retinal diseases but also serves as a general approach for obtaining global information from dense datasets.
- The proposed framework has achieved state-of-the-art results on three retina disease datasets. Additionally, we conducted comprehensive experiments to validate the effectiveness of the proposed components.

Related Works

CNN and Transformer-based Methods. In the field of retinal image analysis, recent research has highlighted the immense potential of deep learning techniques for the automatic recognition and classification of retinal diseases. CNNs and Transformers, known for their robust feature extraction capabilities, have achieved significant success in the automatic diagnosis of retinal diseases (Li et al. 2019; Yu et al. 2021; Zhou et al. 2023; Wang et al. 2023a). However, retinal diseases often present with pathological abnormal regions in fundus images that are small and scattered across the retina. The aforementioned methods lack an inductive bias tailored to this characteristic, resulting in a diminished perception of the pathological visual clues of scattered subtle lesions. Recent studies have introduced saliency maps to direct the network's attention toward areas with subtle lesions (Jiang et al. 2022; Demidov et al. 2023). By focusing on small lesion features, saliency methods have significantly improved the diagnosis of retinal diseases. Nevertheless, these approaches overlook another crucial aspect of retinal disease lesions: their spatial distribution patterns. Understanding these patterns is vital for a more precise identification of retinal diseases.

Graph-based Methods. Graph Neural Networks (GNNs) (Scarselli et al. 2008) extend the capabilities of Convolutional Neural Networks (CNNs) and Transformers to handle irregular and dynamic data structures in image processing. In ophthalmology, several approaches (Liu et al. 2020; Duan et al. 2022; Hu et al. 2023) combine CNNs for feature extraction with GNNs to model category relationships, refining classification results. Alternative studies (Salam et al. 2022; Feng et al. 2023; Sundar and Sumathy 2023; Lei et al. 2024) employ GNNs to capture relationships between features of different diabetic retinopathy grades. However, directly using CNN-extracted or global features in these approaches may lack saliency perception of lesions, hindering the GNN's ability to capture crucial topological information of subtle lesion areas. Effectively modeling the distribution patterns of subtle lesion areas to enhance visual methods for retinal disease diagnosis remains an ongoing challenge.

Methods

As illustrated in Figure 1, the architecture of SIGraph consists of multiple stages, each beginning with a saliency perception block that generates image patch embeddings encapsulating lesion-related saliency information. These embeddings, along with the input image, are subsequently processed by a sequence of image-graph blocks to identify spatial distribution patterns of lesion areas. The multi-stage design progressively integrates lesion area distribution information into image patches at varying resolutions. This it-

erative refinement process enhances the detection of salient lesion cues, leading to a more comprehensive understanding of distribution patterns. In the final stage, a linear classifier utilizes features encompassing visual clues and spatial distribution patterns of lesion areas to classify retinal diseases.

Saliency Perception Block

Given an input image $X \in \mathbb{R}^{h\times w\times 3}$, where h and w represent the image height and width, respectively, the saliency perception block generates saliency features of image patches potentially containing latent lesions. The process begins with the block generating salient features for each patch and then identifying patches that may contain lesions. Specifically, the saliency perception block employs two convolution kernels of different sizes (3×3) and (5×5) with a stride of 2 to sample patches in each stage of the SIGraph. In the first stage, two additional convolution kernels of sizes (9×9) and (27×27) are employed to fully extract features from lesion areas of varying shapes and sizes. Following this, patch embeddings with saliency information are generated using a saliency encoder introduced by (Jiang et al. 2022). These patch embeddings are then fed into a discriminative Multi-layer Perceptron (MLP) to obtain the patch embedding $P \in \mathbb{R}^{n\times d}$, where n represents the number of patches containing lesion pixels, and d denotes the embedding dimension. Subsequently, grouped attention mechanisms are applied to both lesion and non-lesion patches at each stage to enhance their distinctive features. Through this process, n latent lesion patches and their corresponding patch embeddings P are identified.

Image-Graph Block

Using the patch embedding P and image X as inputs, we design an image-graph block comprising three modules to capture the distribution patterns of lesions within latent lesion patches. These three modules achieve key objectives: identify pixel lesion nodes, construct a lesion graph, and extract global distribution patterns of lesion areas. Specifically, we utilize the Pixel2LesionNode module to extract potential lesion pixels as nodes of the lesion graph. Subsequently, we develop the Lesion Graph Construction module to construct a connected lesion graph. Finally, we propose the LesionGraph2Vec module to extract the global distribution features of lesion areas. The details of each module are described as follows:

Pixel2LesionNode Module. When utilizing patch embedding to extract the distribution patterns of subtle lesion areas, a significant challenge arises in adequately representing patches with sparse lesion pixels. These fine-grained details are vulnerable to attenuation or complete loss during downsampling processes. To mitigate this issue, we introduce the Pixel2LesionNode module, which operates at the individual pixel level. This module identifies and represents the smallest portions of lesions as lesion nodes, thereby ensuring the preservation of fine-grained lesion information within the SIGraph framework.

As shown in Figure 2, the Pixel2LesionNode module consists of two parts: generating features of pixels in latent lesion patches and detecting lesion pixels using the generated features. In the first part, given the i-th patch embedding $P_i \in \mathbb{R}^{1\times d}$ and the j-th pixel $c_{ij} \in \mathbb{R}^{1\times 3}$ in the i-th patch, we generate a new pixel feature $p_{ij} \in \mathbb{R}^{1\times d}$. Specifically, we compute the features $l_{ij} \in \mathbb{R}^{1\times d}$ of the pixel c_{ij} by summing the pixel's Laplacian features and patch features:

$$l_{ij} = \phi(Laplacian(c_{ij}, o)) + P_i, \tag{1}$$

where $Laplacian(c_{ij}, o)$ represents the o-order neighbor's Laplacian operation (Tai and Yang 2008), and ϕ denotes a linear transformation that aligns the dimensions of Laplacian features with P_i (from 3 to d). Then, we obtain the pixel feature p_{ij} through a weighted sum operation:

$$s = Softmax(GELU(\gamma(l_{ij}))), \tag{2}$$

$$p_{ij} = l_{ij} + s \cdot l_{ij} \tag{3}$$

where s represents the saliency value of the pixel c_{ij}, $GELU$ denotes the activation functions (Hendrycks and Gimpel 2016), and γ represents a linear transformation to map the dimension of l_{ij} to 1. By applying these operations to all pixels, we obtain pixel-level saliency features in latent lesion patches.

In the second part, we use an additional threshold ε on the pixel saliency value s to identify latent lesion pixels and consider each pixel as the initial node of the lesion graph if its saliency value exceeds ε. In this way, we obtain the lesion nodes V and their initial features F.

Lesion Graph Construction Module. We construct a graph on lesion pixels to capture the global distribution patterns of lesion areas. To effectively capture these global distribution patterns, it is important for the lesion graph to have

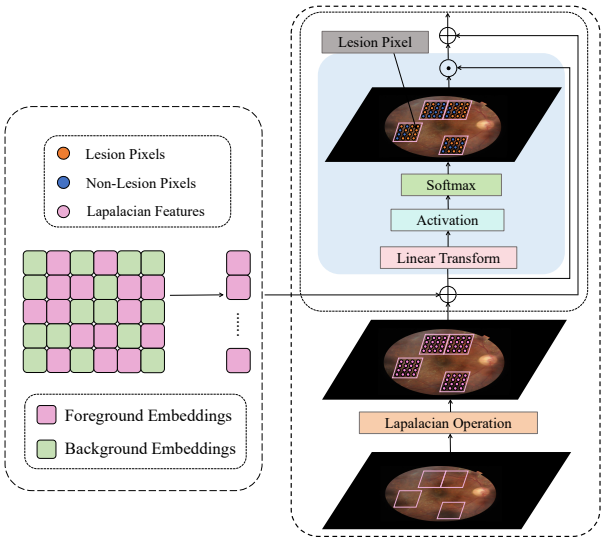

Figure 2: Pixel2LesionNode Module. We calculate the pixel features by combining their Laplacian features with corresponding patch embeddings. Utilizing these pixel features, we generate a saliency map, S, to select potential lesion pixels as nodes for the lesion graph.

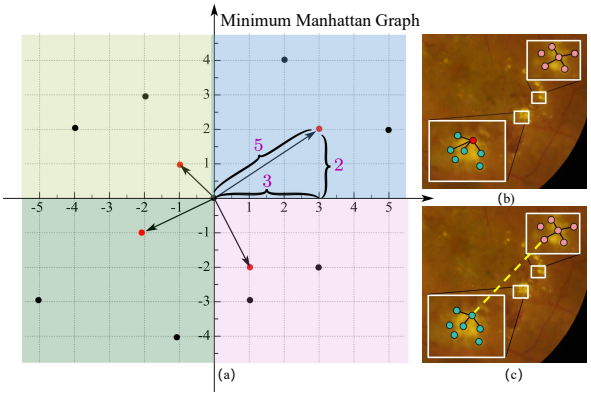

Figure 3: (a) illustrates the selection of adjacent lesion pixels using the minimum Manhattan distance. (b) and (c) shows an example of the connectivity of lesion areas when using the k-NN (k=4) and Manhattan graph construction methods, respectively.

the following two properties: 1) **Sparsity of Edges**. The lesion graph should be sparse, considering lesion areas are typically subtle and scattered. Lesion pixels that are far apart usually belong to different lesion clusters and lack potential relationships. 2) **Full Connectivity**. The lesion graph should be fully connected. If different lesion areas lose connectivity, the resulting distribution will contain only local distribution patterns, leading to suboptimal results. However, commonly used graph construction methods do not meet these two properties. If we simply construct edges between all lesion pixels, the graph network will struggle to extract correct lesion distribution information from the numerous redundant edges. Additionally, using neighborhood-based graph construction methods (e.g., the k-nearest neighbors approach) will lead to disconnected components in the lesion graph (such as the red pixel node in Figure 3(b)).

To address this issue, we propose a new graph construction algorithm based on the minimum Manhattan distance within the four quadrants. As illustrated in Figure 3(a), we redefine the connectivity of each pixel lesion node by splitting the plane associated with each pixel lesion node into four quadrants. Specifically, we build edges between the given pixel lesion node and the nearest pixel lesion nodes in each quadrant. The distance between any two-pixel lesion nodes can be calculated using the Manhattan distance (Singh, Yadav, and Rana 2013). By applying this operation to all pixel lesion nodes, we obtain the edges E of the lesion graph. In addition, for any edge $(v_i, v_j) \in E$, we use the Manhattan distance to calculate the edge weight as follows:

$$W(v_i, v_j) = \frac{h + w - ManhattanDistance(v_i, v_j)}{h + w}. \tag{4}$$

In this way, we construct the lesion graph $G = (V, E; W)$. Our algorithm ensures global connectivity of the lesion graph while maintaining a linear relationship between the number of edges and nodes, specifically $O(4n)$, where n is the number of nodes. This approach is particularly efficient

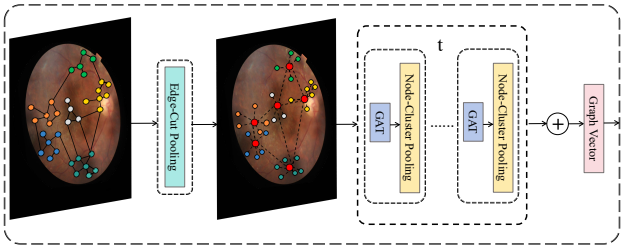

Figure 4: LesionGraph2Vec Module. A two-component mixed pooling operation is used to extract the global distribution patterns.

for constructing connections in dense datasets. The proposed graph construction algorithm effectively addresses the disconnectedness issue illustrated in Figure 3(b). As demonstrated in Figure 3(c), this is achieved by strategically removing an edge in the lower-left direction and adding another in the upper-right direction. This method ensures information exchange between all lesion areas, facilitating a comprehensive analysis of their global distribution patterns.

LesionGraph2Vec Module. To capture the global distribution patterns of lesion areas, a natural method is to utilize message-passing operations (Kipf and Welling 2016; Veličković et al. 2017) on the constructed lesion graph. The main difficulty lies in the graph's massive lesion nodes and sparse edges, which necessitate an extensive number of message-passing operations to effectively extract the global distribution patterns of lesion areas.

Previous studies (Bianchi, Grattarola, and Alippi 2020; Ying et al. 2018; Tsitsulin et al. 2023) have used graph pooling operations to mitigate computational requirements. Nevertheless, the existing pooling operations are not suitable for the lesion graph. For one thing, edge-cut and node-drop pooling operations, although capable of handling large and sparse graphs and efficiently reducing the number of lesion nodes, may lead to the loss of important hierarchical (i.e., the lesion pixels and the lesion cluster to which they belong) and topological information. This loss can ultimately hinder a thorough global understanding. While node-cluster pooling operations can maintain hierarchical and topological information, their edge-folding processes are costly and unstable, especially in large and sparse graphs.

To address these challenges, we propose a new mixed graph pooling operation that can effectively extract global distribution patterns of lesion areas. As illustrated in Figure 4, our mixed pooling operation comprises two components. In the first component, we employ a variant of edge-cut pooling to sever the edges on the boundaries of lesion areas, effectively combining homogeneous and adjacent lesion pixels into clusters. This operation significantly reduces the number of homogeneous lesion pixels, resulting in fewer, more manageable lesion clusters. Subsequently, we apply a variant of node-cluster pooling to the graph of lesion clusters to extract the global distribution patterns of lesion areas with limited computational resources. By introducing our mixed pooling operations, we can use several message-passing op-

erations to derive the global distribution of lesion areas and avoid the loss of hierarchical and topological information. This approach mitigates the disadvantages associated with single-pooling operations. The components of our pooling operation are described as follows:

Edge-cut Pooling Component. In the edge-cut pooling component, we follow a three-step process to reduce the number of lesion pixel nodes to cluster nodes. First, given the lesion graph $G = (V, E; W)$ and the pixel node features F, we calculate the attention weight A_{ij} of each edge $(v_i, v_j) \in E$ as follows:

$$A_{ij} = \frac{\exp(LeakyReLU(\sigma(F_i) \oplus \sigma(F_j)))}{\sum_{k \in \text{adj}(i)} \exp(LeakyReLU(\sigma(F_i) \oplus \sigma(F_k)))}, \tag{5}$$

where F_i and F_j are the pixel lesion node features of v_i and v_j, respectively, adj(i) denotes the neighboring nodes of v_i, σ denotes the linear transform that changes the dimension of F from d to $\frac{d}{2}$, $\oplus$ is the concatenation operation, and $LeakyReLU$ is the activation function (Maas et al. 2013).

After that, we derive new node features F' by aggregating the adjacent nodes based on the attention weight A. Given any pixel lesion node v_i, the node feature F'_i is calculated as follows:

$$F'_i = GELU(\sum_{j \in \text{adj}(i)} A_{ij} \cdot W_{ij} \cdot \delta(F_j)), \tag{6}$$

where δ is the linear transform which does not change the dimension of F. We then use a cut pool operation (Bianchi, Grattarola, and Alippi 2020) on the boundaries of lesion areas and generate lesion cluster nodes C via:

$$(C, G', L_{\text{cut}}) = CutPool(F', G), \tag{7}$$

where C represents the features of the new lesion cluster nodes, $G' = (V', E')$ is the new lesion cluster graph with lesion cluster nodes V' and edges E', and L_{cut} is the auxiliary loss (Bianchi, Grattarola, and Alippi 2020) designed to maximize the probability that each lesion pixel is associated with a single lesion cluster. In this way, we obtain potential lesion clusters, thereby reducing the number of nodes in the lesion graph to a manageable range.

Node-cluster Pooling Component. The node-cluster component consists of t paired Graph Attention layers (GATs) (Veličković et al. 2017) and node-cluster pooling layers. For the i-th pair, we input the lesion cluster graph G'_i and lesion cluster node features $C_i \in \mathbb{R}^{n_i \times d_i}$, where n_i and d_i denote the number of lesion cluster nodes and the feature dimensionality at the i-th layer, respectively. We then compute the downsampled lesion cluster graph G'_{i+1} and corresponding lesion cluster features $C_{i+1} \in \mathbb{R}^{n_{i+1} \times d_{i+1}}$ using a node-cluster pooling operation (Ying et al. 2018). The process is formulated as follows:

$$D_i = GAT(C_i, G_i) \tag{8}$$

$$(C_{i+1}, G'_{i+1}, L_{\text{diff}}) = ClusterPool(D_i, G'_i, r), \tag{9}$$

where D_i are the aggregated cluster features with the same dimension as d_i, r is the edge-folding ratio representing the percentage of reduced edges, and L_{diff} is the auxiliary loss (Ying et al. 2018) that regularizes the difference between G'_i and G'_{i+1}. Specifically, when $i = 1$, G'_1, C_1 represent G', C respectively, and $d_{i+1} = d_i$ except in the first and last node-cluster pooling layers where $d_{i+1} = 2d_i$.

After the final node-cluster pooling layer, we obtain the features $C_t \in \mathbb{R}^{n_t \times 4d}$ of lesion distribution patterns. Furthermore, we calculate the average of C_t to represent lesion global distribution patterns. By integrating the above two pooling components, we tailor the operation to the specific characteristics of lesion graphs, thereby circumventing the limitations of single pooling operations.

Experiment

Datasets

To validate the effectiveness of our method, we conducted experiments on three datasets: EyePACS (Cuadros and Bresnick 2009), RFMiD (Pachade et al. 2021), and a private dataset. The EyePACS dataset comprises 35,126 training images, 10,906 validation images, and 42,607 testing images, aiming to classify the severity of diabetic retinopathy on a scale from 1 to 5. The RFMiD dataset contains 1,920 training images, 640 validation images, and 640 testing images, focusing on classifying retinal images as normal or abnormal. Additionally, our private dataset from a local hospital contains images of various retinal diseases, including 7,032 images of age-related macular degeneration, 6,021 of high myopia, 5,687 of glaucoma, 7,841 of vein occlusion, 10,257 of diabetic retinopathy, and 9,832 normal images. We randomly divided 70% of the dataset for training, 10% for validation, and the remaining 20% for testing. For the binary classification dataset RFMiD, we use AUROC, AUCPR, and F1 metrics as evaluation metrics. We use AUROC, AUC-PR, weighted F1, and weighted Kappa as evaluation metrics for the other two multi-class classification datasets.

Implementation Detail

We initialize our saliency perception block using a pretrained Satformer. As we gradually inject lesion distribution information into the network, we train all parameters of both the saliency perception block and the image-graph block simultaneously. For all images, we use AutoMorph (Zhou et al. 2022) to perform center cropping to remove the background and retain the retina area, resizing them to a resolution of 224 $\times$ 224 as input to SIGraph. We use categorical cross-entropy loss, edge-cut pooling loss L_{cut}, and node-cluster pooling loss L_{diff} with respective weights of 1, 0.2, and 0.1. For quantitative evaluation, we train both the Saliency Perception block and the Image-Graph block with different learning rates of 10^{-4} and 10^{-3}, respectively, using the Adam optimizer for 1000 epochs. It takes around two days to converge on four Nvidia RTX 3090 GPUs.

Results

We compared our method with state-of-the-art methods. The results are presented in Table 1, where SIGraph notably surpasses these advanced architectures in performance. For all

Method Type	Network	Collected Dataset				EyePACS				RFMiD		
		Accuracy	AUC	F1-score	Kappa	Accuracy	AUC	F1-score	Kappa	Accuracy	AUC	F1-score
CNN and GCN methods	Zoom-in-Net (Wang et al. 2017)	79.5	87.4	80.6	83.9	80.7	89.9	78.2	85.4	87.4	90.5	88.9
	GREEN-ResNet50 (Liu et al. 2020)	82.0	89.4	82.2	85.8	82.3	91.4	80.5	87.5	89.6	92.1	91.3
	GRADING (Feng et al. 2023)	82.2	89.5	82.3	86.0	82.4	91.6	80.8	87.6	89.8	92.3	91.4
Transformer Methods	ViT-B/16 (Dosovitskiy et al. 2020)	83.4	92.5	81.3	87.7	83.5	94.1	82.2	87.7	87.1	92.4	90.3
	Swin-B (Liu et al. 2021)	84.1	93.0	82.5	88.6	84.6	95.6	83.5	88.2	88.3	93.2	91.1
	PVTv2-B5 (Wang et al. 2022)	83.0	91.8	81.2	87.3	82.9	93.4	81.3	85.9	87.8	92.4	90.4
	CrossFormer-L (Wang et al. 2023b)	83.5	91.9	81.4	87.7	83.8	93.7	82.4	87.0	90.6	94.3	92.6
	MIL-VT (Yu et al. 2021)	84.2	92.1	81.7	87.9	84.2	94.1	83.3	87.9	91.7	94.8	93.4
	SatFormer-S (Jiang et al. 2022)	86.2	93.8	83.6	89.2	87.4	97.0	86.6	89.8	92.5	95.1	94.3
	SatFormer-B (Jiang et al. 2022)	88.3	94.9	84.8	90.7	88.9	97.7	87.5	90.8	93.8	96.5	95.8
	RETFound (Zhou et al. 2023)	88.5	95.1	85.1	90.8	89.2	97.9	87.8	91.2	93.9	96.6	96.0
Ours	**Saliency Image-Graph**	**91.4**	**98.1**	**88.2**	**93.2**	**90.1**	**98.5**	**88.6**	**92.7**	**95.4**	**98.7**	**98.0**

Table 1: Comparison with state-of-the-art methods on our collected dataset and two benchmark datasets.

Model	Input Size	#Param.(M)	FLOPs(G)
Swin-B	384×384	88	47
ViT-B/16	384×384	86	55
RETFound	256×256	304	190.2
CrossFormer-L	224×224	92	16.1
SatFormer-B	224×224	78	65
SIGraph	224×224	98	104

Table 2: Comparison of parameter amount and computational efficiency with other state-of-the-art methods.

compared models, we maintained their original hyperparameter settings while standardizing the training epochs and data splits to align with SIGraph, ensuring fair comparison when original methods lacked performance metrics on specific datasets.

Comparisons with CNN- and GCN-based Methods. Compared to state-of-the-art CNN- and GCN-based methods, such as Zoom-in-Net, GREEN, and GRADING, our SIGraph achieves a significant performance increase, with at least a 3.1% improvement in Kappa on the EyePACS dataset and a 4.2% increase in F1 score on the RFMiD dataset. Zoom-in-Net, which also utilizes saliency maps, does not account for the interconnections among lesion areas, potentially leading to suboptimal classification results. Conversely, GREEN and GRADING utilize GNNs to establish topological relationships between features derived from CNN. However, they either opt to use the global image feature of the CNN or directly utilize features from different CNN layers, which results in either a loss of fine-grained lesion information or a lack of a saliency map for the lesions. This limitation poses a challenge for GNNs in constructing meaningful relationships between these features. The effectiveness of the SIGraph demonstrates that capturing both fine-grained lesion information and spatial distribution patterns of lesion areas through the lesion graph can significantly enhance classification results.

Comparisons with Transformer-based Methods. Compared to state-of-the-art Transformer-based methods, including Vit-B/16, CrossFormer-L, Swin-B, PVTv2-B5, MIL-VT, SatFormer-S, SatFormer-B, and RETFound, our method demonstrates significant improvements, achieving at least a 1.5% increase in Kappa on EyePACS and a 2.0% increase in F1 Score on RFMiD. Table 2 compares our model's parameters and computational efficiency with those of other models, showing that SIGraph achieves performance enhancements with significantly fewer parameters compared to RETFound. This suggests that by incorporating the image-graph block, our model is able to capture information, specifically the spatial distribution patterns of lesion areas, that large-scale transformers may overlook. When compared to the strong saliency-based transformer method Satformer-B, our SIGraph attains a performance gain of 1.9% in Kappa on EyePACS and 2.2% in F1 Score on RFMiD while maintaining minimal computational costs. This demonstrates that the distribution patterns extracted by the proposed image-graph block can further enhance the model's performance.

Ablation Study

Effectiveness of Image-Graph Block. To evaluate the image-graph block as a plug-and-play component for capturing distribution information on visual features, we applied it to various visual encoders and assessed its effect on overall performance. As the results demonstrated in Table 3, all Transformer-based methods, upon incorporating the image-graph block, achieved at least a 0.8% increase in Kappa and a 1.3% improvement in F1 Score on the EyePACS and RFMiD datasets, respectively, without significantly increasing the number of parameters. This confirms that our image-graph block can generally enhance the performance of transformer frameworks. However, it is important to note that adding the image-graph block to the Swin architecture resulted in a lesser performance boost. This outcome may be attributable to Swin's window-based attention mechanism, which confines updates of visual features within a specific window. Consequently, this mechanism somewhat restricts global perception at each stage. Unlike the Swin Transformer, other transformer methods listed in Table 3 compute attention across the entire image. As a result, the image-graph block can extract a more comprehensive global distribution pattern from the visual clues.

Models	Image-graph Block	#Param.(M)	FLOPs (G)	EyePACS Kappa	RFMiD F1-score
Swin-B	×	88	47	88.2	91.1
	✓	102	86	**89.0**	**92.4**
ViT-B	×	86	55	87.7	90.3
	✓	100	103	**89.0**	**91.9**
MIT-VT	×	98	217	87.9	93.4
	✓	112	265	**89.3**	**95.2**
Satformer-B	×	78	65	90.8	95.8
	✓	98	104	**92.7**	**98.0**

Table 3: Ablation study of adding image-graph block on different transformer methods.

Pooling Methods	Final Clusters	FLOPs (G)	EyePACS Kappa	RFMiD F1-score
Edge-cut	56	79	91.0	96.1
Node-Cluster	56	397	92.6	97.7
Mixed-A	56	96	**92.7**	**98.0**
Mixed-B	114	78	92.3	97.4
Mixed-C	28	102	92.5	97.5

Table 4: Ablation study of different pooling operations on the two public benchmarks.

Model	KNN-4	KNN-8	Manhattan-4	Manhattan-8	EyePACS Kappa	RFMiD F1-score
SIGraph	✓				90.6	95.5
		✓			91.2	96.1
			✓		92.7	97.9
				✓	**92.7**	**98.0**

Table 5: Ablation study of the different graph construction algorithms.

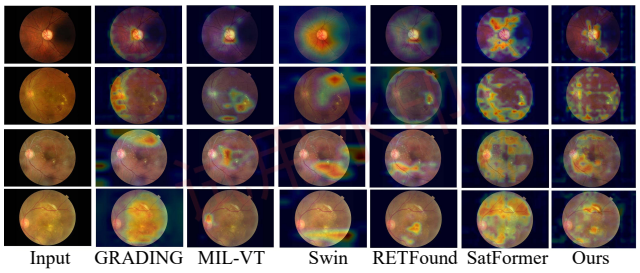

Figure 5: Visualization of the attention maps on the RFMiD test dataset.

Efficacy of Mixed Graph Pooling Operation. We compared different pooling operations and numbers of clusters in the final pooling layer to assess their impact on the model's performance. As depicted in Table 4, when the final clusters were fixed at 56, the edge-cut pooling exhibited the lowest performance among all pooling operations, yielding only marginal improvements of 0.2% in Kappa and 0.4% in F1 score compared to the baseline model Satformer-B. This observation corroborates that excessive edge-cutting operations can result in the loss of crucial hierarchical and topological information pertaining to lesion distributions.

Conversely, while node-based clustering pooling demonstrated accuracy advantages, it incurred substantial computational overhead, particularly when the granularity of lesion nodes approached pixel-level resolution. Our proposed mixed pooling operation, designed to accommodate the subtle and dispersed nature of lesion areas, preserves more information in the lesion graph while significantly reducing computational complexity, thus achieving superior performance among all pooling operations. Furthermore, our experiments with varying the number of nodes in the final pooling layer revealed that approximately 50 nodes as the final node clusters achieved optimal balance. Both increasing the number of nodes and further pooling to compress node information led to degradation in model performance.

Importance of Graph Construction Algorithm. Table 5 illustrates the influence of various lesion graph construction methods on the classification performance. The results obtained from the EyePACS and RFMiD datasets demonstrate that lesion graphs constructed using our proposed method achieved significant performance enhancements, with improvements of at least 1.5% in Kappa and 1.8% in F1 score, respectively. Notably, the four-neighbor k-NN-based graph construction method resulted in a performance degradation. This decline can be attributed to the potential disconnection of the lesion graph, causing the network to overly focus on isolated lesion areas. Furthermore, we conducted an ablation study to evaluate the impact of the number of neighbors selected in each quadrant in our Manhattan method. The results indicate that selecting a single node per quadrant is sufficient for effective information exchange between different lesion areas. Increasing the neighboring nodes yielded diminishing returns relative to the additional computational overhead incurred.

Visualization of Attention Maps

We provided visualizations to compare how different methods guide the identification of lesion areas for disease diagnosis in Figure 5. We observed that non-saliency methods, which lack saliency maps and rely solely on convolution and attention mechanisms, often excessively focus on the most salient features of the image, thereby losing perception of subtle lesion areas. Methods that use saliency maps demonstrate a notable improvement in the granularity of focus areas compared to other approaches. However, relying solely on a saliency map is insufficient for accurately identifying lesion areas, often resulting in a focus on incorrect saliency areas, especially in non-lesion areas close to actual lesions. In contrast, SIGraph prioritizes areas with finer granularity while simultaneously perceiving more accurate lesion areas.

Conclusions

In this paper, we present the Saliency Image-Graph network, a novel network employing saliency features. This method augments saliency lesion features by integrating the spatial distribution patterns of lesion areas. We construct a fully connected graph based on the lesion pixels and propose a mixed graph pooling operation to extract the global distribution patterns of these lesion areas effectively. Experimental results demonstrate that SIGraph, by combining visual and distribution features of lesions, achieves state-of-the-art performance in diagnosing retinal diseases.

Acknowledgments

This research was partially supported by the Key Research and Development Program of Zhejiang Province under Grant 2021C03032.

References

Bianchi, F. M.; Grattarola, D.; and Alippi, C. 2020. Spectral clustering with graph neural networks for graph pooling. In *International conference on machine learning*, 874–883.

Cuadros, J.; and Bresnick, G. 2009. EyePACS: an adaptable telemedicine system for diabetic retinopathy screening. *Journal of diabetes science and technology*, 3(3): 509–516.

Demidov, D.; Sharif, M. H.; Abdurahimov, A.; Cholakkal, H.; and Khan, F. S. 2023. Salient Mask-Guided Vision Transformer for Fine-Grained Classification. *arXiv preprint arXiv:2305.07102*.

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*.

Duan, S.; Huang, P.; Chen, M.; Wang, T.; Sun, X.; Chen, M.; Dong, X.; Jiang, Z.; and Li, D. 2022. Semi-supervised classification of fundus images combined with CNN and GCN. *Journal of Applied Clinical Medical Physics*, 23(12): e13746.

Feng, M.; Wang, J.; Wen, K.; and Sun, J. 2023. Grading of Diabetic Retinopathy Images Based on Graph Neural Network. *IEEE Access*.

Han, D.; Yun, S.; Heo, B.; and Yoo, Y. 2021. Rethinking channel dimensions for efficient model design. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 732–741.

Hendrycks, D.; and Gimpel, K. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Hu, M.; Wang, J.; Wynne, J.; Liu, T.; and Yang, X. 2023. A vision-GNN framework for retinopathy classification using optical coherence tomography. In *Medical Imaging 2023: Computer-Aided Diagnosis*, volume 12465, 200–206. SPIE.

Jiang, Y.; Xu, K.; Wang, X.; Li, Y.; Cui, H.; Tao, Y.; and Lin, H. 2022. SatFormer: Saliency-Guided Abnormality-Aware Transformer for Retinal Disease Classification in Fundus Image. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, 987–994.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Lei, Y.; Lin, S.; Li, Z.; Zhang, Y.; and Lai, T. 2024. GNN-fused CapsNet with multi-head prediction for diabetic retinopathy grading. *Engineering Applications of Artificial Intelligence*, 133: 107994.

Li, X.; Hu, X.; Yu, L.; Zhu, L.; Fu, C.-W.; and Heng, P.-A. 2019. CANet: cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading. *IEEE transactions on medical imaging*, 39(5): 1483–1493.

Liu, S.; Gong, L.; Ma, K.; and Zheng, Y. 2020. GREEN: a graph residual re-ranking network for grading diabetic retinopathy. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020*, 585–594.

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.

Maas, A. L.; Hannun, A. Y.; Ng, A. Y.; et al. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, 3.

Ong, A. Y.; Birtel, J.; Agorogiannis, E.; Sharma, S. M.; and Charbel Issa, P. 2023. Topographic patterns of retinal lesions in multiple evanescent white dot syndrome. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 1–8.

Pachade, S.; Porwal, P.; Thulkar, D.; Kokare, M.; Deshmukh, G.; Sahasrabuddhe, V.; Giancardo, L.; Quellec, G.; and Mériaudeau, F. 2021. Retinal fundus multi-disease image dataset (RFMiD): A dataset for multi-disease detection research. *Data*, 6(2): 14.

Salam, A. A.; Mahadevappa, M.; Das, A.; and Nair, M. S. 2022. DRG-NET: A graph neural network for computer-aided grading of diabetic retinopathy. *Signal, Image and Video Processing*, 16(7): 1869–1875.

Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1): 61–80.

Singh, A.; Yadav, A.; and Rana, A. 2013. K-means with Three different Distance Metrics. *International Journal of Computer Applications*, 67(10).

Sundar, S.; and Sumathy, S. 2023. Classification of Diabetic Retinopathy disease levels by extracting topological features using Graph Neural Networks. *IEEE Access*, 11: 51435–51444.

Tai, S.-C.; and Yang, S.-M. 2008. A fast method for image noise estimation using laplacian operator and adaptive edge detection. In *2008 3rd International Symposium on Communications, Control and Signal Processing*, 1077–1081.

Tsitsulin, A.; Palowitch, J.; Perozzi, B.; and Müller, E. 2023. Graph clustering with graph neural networks. *Journal of Machine Learning Research*, 24(127): 1–21.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Walter, T.; Klein, J.-C.; Massin, P.; and Erginay, A. 2002. A contribution of image processing to the diagnosis of diabetic retinopathy-detection of exudates in color fundus images of the human retina. *IEEE transactions on medical imaging*, 21(10): 1236–1243.

Wang, K.; Xu, C.; Li, G.; Zhang, Y.; Zheng, Y.; and Sun, C. 2023a. Combining convolutional neural networks and self-attention for fundus diseases identification. *Scientific Reports*, 13(1): 76.

Wang, W.; Chen, W.; Qiu, Q.; Chen, L.; Wu, B.; Lin, B.; He, X.; and Liu, W. 2023b. Crossformer++: A versatile vision

transformer hinging on cross-scale attention. *arXiv preprint arXiv:2303.06908*.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2022. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3): 415–424.

Wang, Z.; Yin, Y.; Shi, J.; Fang, W.; Li, H.; and Wang, X. 2017. Zoom-in-net: Deep mining lesions for diabetic retinopathy detection. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017*, 267–275.

Ying, Z.; You, J.; Morris, C.; Ren, X.; Hamilton, W.; and Leskovec, J. 2018. Hierarchical graph representation learning with differentiable pooling. *Advances in neural information processing systems*, 31.

Yu, S.; Ma, K.; Bi, Q.; Bian, C.; Ning, M.; He, N.; Li, Y.; Liu, H.; and Zheng, Y. 2021. Mil-vt: Multiple instance learning enhanced vision transformer for fundus image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021*, 45–54. Springer.

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*, 1–8.

Zhou, Y.; Wagner, S. K.; Chia, M. A.; Zhao, A.; Xu, M.; Struyven, R.; Alexander, D. C.; Keane, P. A.; et al. 2022. AutoMorph: automated retinal vascular morphology quantification via a deep learning pipeline. *Translational vision science & technology*, 11(7): 12–12.