Frequency-Aligned Cross-Modal Learning with Top-K Wavelet Fusion and Dynamic Expert Routing for Enhanced Retinal Disease Diagnosis

Yuxin Lin[1,2], Haoran Li[3], Haoyu Cao[1], Yongting Hu[1], Qihao Xu[1], Chengliang Liu[4], Xiaoling Luo[5], Zhihao Wu[6], Yong Xu[1,2], Wei Wang[1] *

[1] School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen
[2] Shenzhen Key Laboratory of Visual Object Detection and Recognition
[3] School of Information Technology, University of Wollongong
[4] Laboratory for Artificial Intelligence in Design, The Hong Kong Polytechnic University
[5] College of Computer Science and Software Engineering, Shenzhen University
[6] School of Artificial Intelligence, Shenzhen University
linyuxin6688@gmail.com, hl644@uowmail.edu.au, haoyucao1016@gmail.com, huyongting08@163.com, xqh51199597@outlook.com, liucl1996@163.com, xiaolingluoo@outlook.com, horatio_ng@163.com, laterfall@hit.edu.cn, wangwei2019@hit.edu.cn

Abstract

Multimodal fusion of color fundus photography (CFP) and optical coherence tomography (OCT) B-scan images has demonstrated superior diagnostic potential for retinal diseases compared to single-modality approaches. However, existing fusion paradigms - whether through naive concatenation or attention mechanisms - treat cross-modal interactions indiscriminately, lacking adaptive modulation of modality-specific contributions under varying clinical scenarios. We propose an adaptive fusion framework that dynamically routes and refines multimodal signals for enhancing disease recognition. The framework comprises two key components: 1) Dynamic Cross-Modal Expert Routing (CMER), which selectively activates convolutional neural network (CNN) experts from one modality based on contextual guidance from the other, ensuring only the most relevant feature extractors contribute to fusion; and 2) Top-K Expert-Guided Wavelet Fusion (TEWF), which performs discrete wavelet transform (DWT) to decompose selected features into low- and high-frequency subbands. Cross-modal attention is then applied specifically to high-frequency components, where lesion-specific microstructures reside, enabling frequency-aware fusion. Finally, inverse DWT (IDWT) reconstructs the fused representation, weighted by CMER-derived importance scores to amplify informative modality cues while suppressing redundancy. Experimental validation on two multimodal retinal datasets demonstrates that our method achieves state-of-the-art performance, outperforming existing fusion strategies by significant margins in disease classification accuracy and robustness.

Introduction

According to a recent report by the World Health Organization (WHO), approximately 216.6 million people worldwide suffer from moderate to severe vision impairment, with 36 million individuals classified as blind (Lin et al. 2025b). Epidemiological data indicate that at least 45% of such vision impairments can be fully reversed if detected early and managed promptly (Ferris and Tielsch 2004; Luo et al. 2025b). The principal fundus diseases leading to irreversible vision loss include diabetic retinopathy (DR) , glaucoma, cataract, myopic retinopathy (MR), and age-related macular degeneration (AMD)(Luo et al. 2024). Early detection is critical for these conditions, since therapeutic efficacy declines sharply once lesions advance beyond a controllable stage (Bilal, Keles, and Bendechache 2025). Current screening programs for high-risk populations are constrained by uneven distribution of medical resources and a shortage of qualified professionals, challenges that are most severe in remote or resource-limited settings (Wen et al. 2025; Luo et al. 2025a). Manual fundus examination is time-consuming, labor-intensive, and subject to inter-observer variability, making it difficult to scale in order to meet rising diagnostic demand (Lin et al.

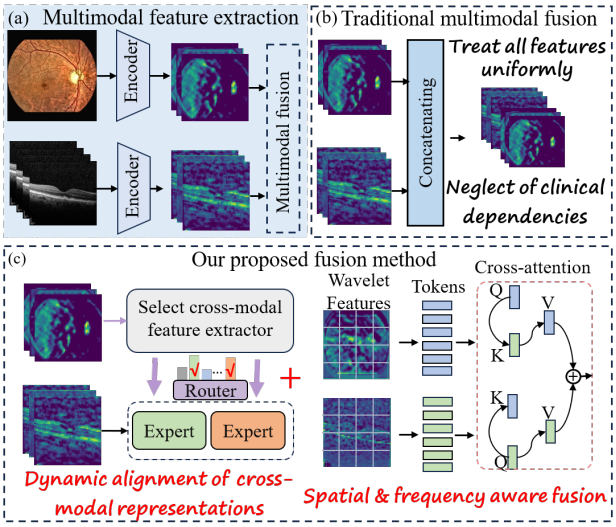


Figure 1: Comparison of traditional fusion methods and our proposed method

*This author is the corresponding author.

2025c).

Deep learning has shown promise in automated retinal diagnosis using CFP and OCT modalities. CFP provides a high spatial resolution facial retinal surface visualization, capturing vascular abnormalities and pigmentary changes (Lin et al. 2025b), while OCT delivers cross-sectional tomograms revealing subsurface microstructures such as intraretinal fluid and hyperreflective foci (Li et al. 2024). Single-modality approaches inherently miss complementary pathological cues: CFP may overlook sub-retinal fluid accumulation, while OCT could fail to characterize diffuse hemorrhages or neovascular patterns (Wang et al. 2019). In recent years, multimodal learning has shown remarkable progress in deep learning, effectively overcoming the inherent limitations of single-modal approaches (Zhang et al. 2024c,b; Qin, Feng, and Zhang 2025), which also provides valuable inspiration for fundus image analysis (Lei et al. 2024). Multimodal fusion (Figure. 1(a)) has thus emerged as a paradigm shift, yet existing methods rely on undifferentiated fusion operations - such as feature concatenation or uniform attention - that treat all cross-modal interactions equivalently (Figure. 1(b)). This approach neglects clinically meaningful modality dependencies, as different imaging modalities contribute complementary information at varying spatial scales and frequency bands.

To address these limitations, we propose a two-stage adaptive fusion framework (Figure 1(c)) that explicitly incorporates modality-specific dependencies. The first stage, the CMER module, introduces modality-specific CNN expert banks, where the activation of experts in one modality is guided by contextual signals from the other. This selective routing ensures the alignment of features that reflect the unique contributions of each modality, capturing modality dependencies and preserving the complementary nature of the different data sources. The second stage, the TEWF module, further enhances this by decomposing features into wavelet subbands using DWT. Cross-modal attention is applied specifically to high-frequency components, where critical pathological details—such as lesions—are concentrated. Subsequently, the final fused representation is reconstructed by combining the preserved low-frequency anatomical structures with these refined, attention-enhanced high-frequency pathological cues through IDWT. This dual spatial-frequency fusion enables our method to preserve coarse anatomical structures (low-frequency) while emphasizing fine-grained pathological cues (high-frequency), effectively addressing the varied contributions of different modalities. The contributions of this paper are as follows:

(1) A CMER mechanism that achieves context-aware modality alignment by dynamically routing CNN experts based on cross-modal guidance, moving beyond traditional feature fusion.

(2) We propose a TEWF strategy that combines frequency alignment with spatial-domain fusion on wavelet-transformed high-frequency components, explicitly fusing cross-modal salient microstructures.

(3) Extensive experiments on two multimodal retinal datasets show that our framework consistently outperforms existing fusion approaches, achieving the highest disease recognition accuracy and robustness.

Related Works

Single Modal Retinal Disease Recognition

Single-modality approaches to retinal disease recognition have primarily focused on either CFP or OCT in isolation. For CFP-based methods, researchers leverage high-resolution fundus images to analyze lesion morphology, vascular abnormalities, and optic disc changes. Zhang et al. proposed SIGraph, which constructs a graph over subtle lesion regions to capture spatial distribution patterns, achieving notable improvements in diagnostic accuracy (Zhang et al. 2025). Wen et al. introduced a lesion-guided transformer architecture that models long-range dependencies among heterogeneous lesion types, demonstrating superior performance in complex disease recognition (Wen et al. 2024). To address the limited field-of-view limitation in single-field CFP, Lin et al. developed a multi-view attention network that models inter-view focus relationships, significantly enhancing DR detection accuracy (Lin et al. 2025a). Su et al. tackled domain adaptation challenges in clinical settings through generative unadversarial examples (GUES), specifically designed for online model-diagnostic adaptation (Su et al. 2025).

For OCT-based approaches, methods exploit volumetric B-scans to reveal subsurface microstructural alterations. Yang et al. designed a weakly supervised segmentation pipeline that localizes and classifies anomalies at the slice level using only image-level labels, facilitating early lesion detection (Yang et al. 2024). Zhang et al. presented a regional context-based recalibration module that incorporates clinical priors through region pooling and context fusion, substantially improving cataract recognition accuracy (Zhang et al. 2024a).

Multi Modal Retinal Disease Recognition

Recent advances in deep learning have exposed performance bottlenecks in unimodal medical image analysis, spurring increased interest in multimodal fusion strategies (Yu et al. 2019). Notably, Yoo et al. pioneered the exploration of combining OCT and CFP modalities for AMD diagnosis, demonstrating that multimodal deep learning frameworks could surpass unimodal baselines (Yoo et al. 2019). Subsequent work by Wang et al. advanced this paradigm by employing two-stream CNN architectures to separately process OCT and CFP data before late fusion via concatenation, achieving state-of-the-art AMD classification accuracy (Wang et al. 2022). More recently, Zou et al. introduced bayesian uncertainty modeling for glaucoma detection (Zou et al. 2024), using normal-inverse-gamma priors to quantify modality-specific uncertainties prior to feature-level concatenation. However the above methods typically employ naive concatenation or homogeneous attention mechanisms that treat cross-modal interactions equivalently, failing to adaptively modulate modality contributions based on lesion characteristics or clinical context. Recent work by Wang et al. has introduced a CLIP-inspired contrastive learning framework for OCT-CFP multimodal analysis, achieving

great performance in retinal disease recognition (Wang et al. 2024). By aligning OCT volumes and CFP images through a shared projection space with cross-modal contrastive loss, this method effectively bridges the representation gap between OCT and CFP data.

The Proposed Method

Figure 2 provides an overview of the proposed model, which builds upon the dual-branches ResNet18 architecture, widely used in multi-modal retinal disease recognition tasks (Wang et al. 2022). As illustrated in Figure 2(a), the initial feature extraction for both modalities is performed through the network's initial layer, adhering to the standard ResNet18 structure. Following this, the processing pipeline progresses through four distinct stages. Each stage leverages the CNN-based layers (layer 1 to layer 4) of ResNet18 to extract modality-specific features. At each stage, these features serve three primary functions: First, they act as the current modality's feature representation, which is subsequently fused with the features from the other modality. Second, they provide routing signals (indicated by the purple arrow in Figure. 2) to guide the selection of the most relevant CNN experts from the other modality. Finally, these features are used as the source for fusion (depicted by the light red arrow in Figure. 2), contributing to the final fused representation of the other modality. This process not only ensures efficient feature extraction within each modality's ResNet18 stream, but also enhances disease recognition accuracy by enabling cross-modal feature fusion through the proposed CMER and TEWF frameworks.

In this paper, we will distinguish between the different branches inputs (the OCT and the CFP) by labeling the variable names and module names in the OCT branch with superscripts $*^O$, and those in the CFP branch with $*^C$. Additionally, variables and modules at different stages will be labeled with the subscript $*_{s_i}$, where s_i represents the i-th stage. For example, $\mathrm{CMER}^O_{s_1}$ refers to the CMER module in stage 1 of the OCT branch.

Dynamically Cross-Modal Expert Routing

Let $\boldsymbol{\mathcal{F}}^O_{s_i}$ and $\boldsymbol{\mathcal{F}}^C_{s_i}$ denote the feature representations obtained from layer i of the OCT and CFP processing branches, respectively. As illustrated in Figure 2(b), we showcase the mechanism for cross-modal feature extraction by dynamically selecting OCT CNN experts based on complementary CFP modality information. Importantly, our proposed framework facilitates mutual cross-modal collaboration; thus, a similar cross-modal integration, in which OCT features guide the selection of CFP CNN experts, is also performed. This bidirectional cross-modal interaction ensures balanced and effective feature extraction from both OCT and CFP modalities.

Specifically, we implement this adaptive selection mechanism through our CMER module. Taking CFP-to-OCT feature extraction as an illustrative example, we first process CFP features $\boldsymbol{\mathcal{F}}^C_{s_i}$ via a routing network designed to determine optimal OCT expert activation. Spatial patterns in $\boldsymbol{\mathcal{F}}^C_{s_i}$ are encoded using a CNN block with adaptive average pooling (denoted as CB_a) which preserves translational invariance and reduces dimensionality. The resultant feature map is flattened and projected through a linear transformation (denoted as Lr) to yield routing logits. Then, these logits are normalized by a softmax operation into routing scores:

$$\boldsymbol{\mathcal{R}s}^O_{s_i} = \mathrm{Softmax}\left(\mathrm{Lr}\left(\mathrm{CB}_a(\boldsymbol{\mathcal{F}}^C_{s_i})\right)\right) \in R^N \quad (1)$$

where N denotes the number of pre-defined OCT experts. Inspired by the Mixture-of-Experts (MoE) framework (Cao et al. 2023), the OCT features $\boldsymbol{\mathcal{F}}^O_{s_i}$ are fed into a set of activated OCT CNN experts $\left\{\ \mathrm{CE}^O_{s_i,1}\ \ \mathrm{CE}^O_{s_i,2}\ \ \cdots\ \ \mathrm{CE}^O_{s_i,K}\ \right\}$. Each expert $\mathrm{CE}^O_{s_i,k}$ is selectively activated according to the ranking of values within the routing scores $\boldsymbol{\mathcal{R}s}^O_{s_i}$, specifically triggered by the k-th highest score. The dynamically extracted OCT features are computed as:

$$\boldsymbol{\mathcal{F}e}^O_{s_i,k} = \mathrm{CE}^O_{s_i,k}\left(\boldsymbol{\mathcal{F}}^O_{s_i}\right) \quad (2)$$

Subsequently, the extracted OCT feature $\boldsymbol{\mathcal{F}e}^O_{s_i,k}$, along with its corresponding importance weight $\boldsymbol{w}^O_{s_i,k}$, the k-th highest value from $\boldsymbol{\mathcal{R}s}^O_{s_i}$, and the CFP feature $\boldsymbol{\mathcal{F}}^C_{s_i}$ are jointly input into the subsequent $\mathrm{TEWF}^O_{s_i}$ module, facilitating effective and adaptive cross-modality fusion. Similarly, the complementary procedure will generate the extracted CFP features $\boldsymbol{\mathcal{F}e}^C_{s_i,k}$ and their corresponding weights $\boldsymbol{w}^C_{s_i,k}$, which will be processed by the $\mathrm{TEWF}^C_{s_i}$. For convenience, only the OCT branch is illustrated in this paper.

Similar to existing MoE-based methods (Xie et al. 2025), we incorporate a load-balancing loss term ($\boldsymbol{\mathcal{L}}_{load}$) to encourage equitable utilization of CNN experts, ensuring that each expert processes roughly equal numbers of training examples during optimization.

Top-K Expert-Guided Wavelet Fusion

Figure 2(c) illustrates the detailed workflow of our proposed TEWF module, highlighting the multimodal fusion procedure. TEWF consists of three primary components: (1) DWT and IDWT, which decompose multimodal features into distinct spatial-frequency subbands and reconstruct fused wavelet components; (2) a cross-attention module on higher frequency wavelet subbands, which selectively fuses multimodal information across different wavelet subregions, particularly emphasizing lesion-related high-frequency details; and (3) a weighted sum mechanism, leveraging weights derived from the CMER module to further enhance fusion effect by promoting highly relevant cross-modal information while suppressing redundancy.

Specifically, the selectively activated OCT features $\boldsymbol{\mathcal{F}e}^O_{s_i,k}$ and the corresponding CFP modality features $\boldsymbol{\mathcal{F}}^C_{s_i}$ are decomposed into wavelet subbands using the DWT:

$$\boldsymbol{\mathcal{LL}}^C_{s_i}, \boldsymbol{\mathcal{LH}}^C_{s_i}, \boldsymbol{\mathcal{HL}}^C_{s_i}, \boldsymbol{\mathcal{HH}}^C_{s_i} = \mathrm{DWT}\left(\boldsymbol{\mathcal{F}}^C_{s_i}\right) \quad (3)$$

$$\boldsymbol{\mathcal{LL}}^O_{s_i,k}, \boldsymbol{\mathcal{LH}}^O_{s_i,k}, \boldsymbol{\mathcal{HL}}^O_{s_i,k}, \boldsymbol{\mathcal{HH}}^O_{s_i,k} = \mathrm{DWT}\left(\boldsymbol{\mathcal{F}e}^O_{s_i,k}\right) \quad (4)$$

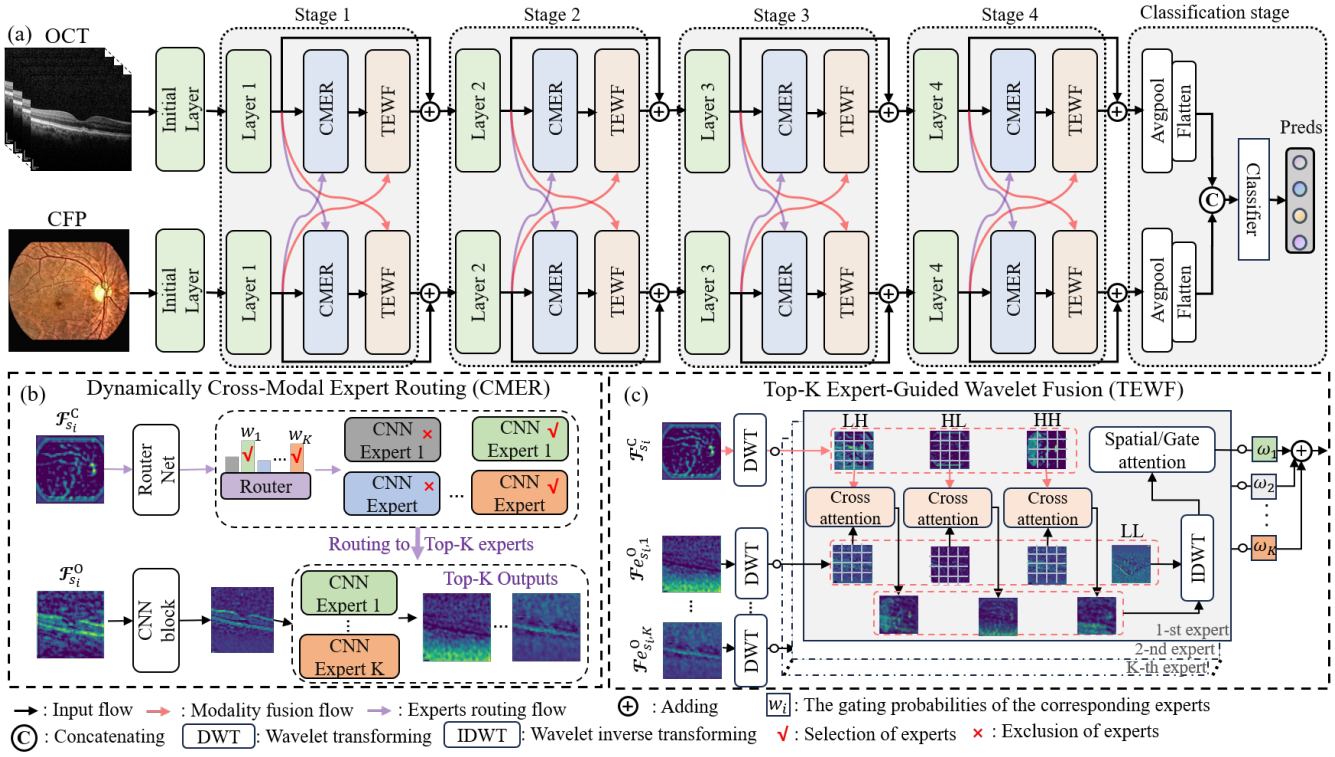


Figure 2: Overview of our proposed model. (a). The workflow of our model; (b) The proposed CMER; (c) The proposed TEWF

where $k = [1, 2, \cdots K]$. The LL subband captures the overall structural and smoothly varying features of each modality, while the LH, HL, and HH subbands represent detailed information along the horizontal, vertical, and diagonal directions, respectively. In retinal imaging, disease-related lesions typically exhibit fine-grained structures characterized by strong edges, which predominantly reside within these high-frequency subbands (LH, HL, and HH) (Cao et al. 2020). Consequently, our multimodal fusion specifically targets these high-frequency wavelet components to effectively integrate clinically significant lesion details from both modalities.

Similar to the tokenizing operation of most Vision Transformer (ViT)-based approaches (Lin et al. 2025b), each high-frequency feature map is partitioned into non-overlapping patches and subsequently tokenized. For example, in the LH subband, we obtain sequential representations $\mathcal{T}_{s_i,k}^{O,LH}$ and $\mathcal{T}_{s_i}^{C,LH}$ for the OCT and CFP modalities, respectively. These tokenized features are then projected into corresponding query matrices, $\left\{\mathcal{Q}_{s_i,k}^{O,LH}, \mathcal{Q}_{s_i,k}^{O,HL}, \mathcal{Q}_{s_i,k}^{O,HH}\right\}$and $\left\{\mathcal{Q}_{s_i}^{C,LH}, \mathcal{Q}_{s_i}^{C,HL}, \mathcal{Q}_{s_i}^{C,HH}\right\}$ for OCT and CFP modalities, respectively. Likewise, corresponding key matrices $\left\{\mathcal{K}_{s_i,k}^{O,LH}, \mathcal{K}_{s_i,k}^{O,HL}, \mathcal{K}_{s_i,k}^{O,HH}\right\}$ and $\left\{\mathcal{K}_{s_i}^{C,LH}, \mathcal{K}_{s_i}^{C,HL}, \mathcal{K}_{s_i}^{C,HH}\right\}$, as well as value matrices $\left\{\mathcal{V}_{s_i,k}^{O,LH}, \mathcal{V}_{s_i,k}^{O,HL}, \mathcal{V}_{s_i,k}^{O,HH}\right\}$ and $\left\{\mathcal{V}_{s_i}^{C,LH}, \mathcal{V}_{s_i}^{C,HL}, \mathcal{V}_{s_i}^{C,HH}\right\}$ are generated for each modality. Then the fusion are perfomed on the same level of the wavelet components across the modalities. Taking the fusion of the LH subband as an example, the cross-attention mechanism is implemented through two complementary stages: OCT-to-CFP and CFP-to-OCT attention, respectively formulated as follows:

$$\begin{aligned}
\boldsymbol{attn}_{s_i,k}^{O,LH,1} &= \mathrm{Lr}\left(\mathrm{Softmax}\left(\frac{\mathcal{Q}_{s_i,k}^{O,LH}\mathcal{K}_{s_i}^{C,LH}}{\sqrt{D}}\right)\mathcal{V}_{s_i,k}^{O,LH}\right) \\
\boldsymbol{attn}_{s_i,k}^{O,LH,2} &= \mathrm{Lr}\left(\mathrm{Softmax}\left(\frac{\mathcal{Q}_{s_i}^{C,LH}\mathcal{K}_{s_i,k}^{O,LH}}{\sqrt{D}}\right)\mathcal{V}_{s_i}^{C,LH}\right)
\end{aligned} \tag{5}$$

Subsequently, these bidirectional attention results are integrated by averaging:

$$\boldsymbol{attn}_{s_i,k}^{O,LH} = \frac{\boldsymbol{attn}_{s_i,k}^{O,LH,1} + \boldsymbol{attn}_{s_i,k}^{O,LH,2}}{2} \tag{6}$$

Finally, the fused features are obtained by adding the residual tokens $\mathcal{T}_{s_i,k}^{O,LH}$ and reshaping (Re) the resulting sequencial tokens back into their original wavelet subband 2D form:

$$\boldsymbol{fu}_{s_i,k}^{O,LH} = \mathrm{Re}\left(\mathcal{T}_{s_i,k}^{O,LH} + \boldsymbol{attn}_{s_i,k}^{O,LH}\right) \tag{7}$$

It is worth noting that the residual addition of $\mathcal{T}_{s_i,k}^{O,LH}$

aligns with standard transformer architectures, typically implemented across multiple layers. Here, for clarity and brevity, only one representative layer is explicitly shown in the equations.

Finally, the fused representation corresponding to the $k-th$ expert is generated by combining the cross-modal fused high-frequency wavelet subbands ($\boldsymbol{fu}^{O,LH}_{s_i,k}$, $\boldsymbol{fu}^{O,HL}_{s_i,k}$, $\boldsymbol{fu}^{O,HH}_{s_i,k}$) with the original low-frequency subband ($\boldsymbol{\mathcal{LL}}^{O}_{s_i,k}$) via IDWT:

$$\tilde{\boldsymbol{\mathcal{F}}}\boldsymbol{e}^{O}_{s_i,k} = \mathrm{IDWT}\left(\boldsymbol{\mathcal{LL}}^{O}_{s_i,k}, \boldsymbol{fu}^{O,LH}_{s_i,k}, \boldsymbol{fu}^{O,HL}_{s_i,k}, \boldsymbol{fu}^{O,HH}_{s_i,k}\right) \quad (8)$$

Furthermore, $\tilde{\boldsymbol{\mathcal{F}}}\boldsymbol{e}^{O}_{s_i,k}$ is processed by applying both gate attention and spatial attention as follows:

$$\hat{\boldsymbol{\mathcal{F}}}\boldsymbol{e}^{O}_{s_i,k} = \tilde{\boldsymbol{\mathcal{F}}}\boldsymbol{e}^{O}_{s_i,k} \times \mathrm{G}(\tilde{\boldsymbol{\mathcal{F}}}\boldsymbol{e}^{O}_{s_i,k}) \times \mathrm{Sp}\left(\tilde{\boldsymbol{\mathcal{F}}}\boldsymbol{e}^{O}_{s_i,k}\right) \quad (9)$$

where G denotes the gate attention mechanism, and Sp represents the spatial attention function. Subsequently, the final output feature from the CMER^{O}_{si} module is obtained by performing a weighted sum over all experts' fused representations, as expressed by:

$$\tilde{\boldsymbol{\mathcal{F}}}^{O}_{s_i} = \sum_{k=1}^{K} \boldsymbol{w}^{O}_{s_i,k} \times \hat{\boldsymbol{\mathcal{F}}}\boldsymbol{e}^{O}_{s_i,k} \quad (10)$$

The proposed TEWF module offers a principled and effective strategy for modality fusion by simultaneously leveraging spatial-frequency decomposition and dynamic expert weighting. Unlike conventional fusion methods that often rely on spatial-domain concatenation or uniform attention, TEWF decomposes modality-specific features into high- and low-frequency subbands using DWT, thereby isolating fine-grained pathological details critical for retinal disease recognition. By focusing fusion on high-frequency components—where lesion-specific signals such as edges and textures are most prominent—our method enhances the sensitivity to subtle disease cues. Furthermore, the bidirectional cross-attention mechanism enables context-aware alignment between modalities at a subregion level, ensuring complementary information is selectively integrated. Finally, through IDWT and expert-guided weighted aggregation, TEWF adaptively reconstructs the fused representation while amplifying relevant modality cues and suppressing redundancy. This design not only improves fusion fidelity but also enhances the overall robustness and interpretability of the multimodal representation.

Classification Stage

As shown in Figure 2(a), the classification stage operates with dual inputs, incorporating the outputs of CMER^{O}_{s4} and the CMER^{C}_{s4}, respectively. Residual connections are applied to enhance gradient flow and preserve original modality information (Liu et al. 2025), formulated as: $\boldsymbol{\mathcal{I}}^{O} = \tilde{\boldsymbol{\mathcal{F}}}^{O}_{s_4} + \boldsymbol{\mathcal{F}}^{O}_{s_4}$ and $\boldsymbol{\mathcal{I}}^{C} = \tilde{\boldsymbol{\mathcal{F}}}^{C}_{s_4} + \boldsymbol{\mathcal{F}}^{C}_{s_4}$. The final prediction is computed as:

$$\boldsymbol{\mathcal{P}} = \mathrm{Lr}\left(\mathrm{Conc}\left(\mathrm{Fl}\left(\mathrm{Avp}\left(\boldsymbol{\mathcal{I}}^{C}\right)\right), \mathrm{Fl}\left(\mathrm{Avp}\left(\boldsymbol{\mathcal{I}}^{O}\right)\right)\right)\right) \quad (11)$$

where $\mathrm{Avp}()$ denotes adaptive average pooling, $\mathrm{Fl}()$ is the flattening performance, $\mathrm{Conc}()$ represents the concatenation. This architecture ensures that modality-specific and fused features are jointly considered in the decision-making process.

The overall training objective integrates the task-specific cross-entropy loss ($\boldsymbol{\mathcal{L}}_{task}$) for classification and the expert load-balancing loss ($\boldsymbol{\mathcal{L}}_{load}$) introduced in the CMER module to ensure uniform expert utilization. The final loss function is dynamically weighted based on the relative magnitude of $\boldsymbol{\mathcal{L}}_{load}$, as follows:

$$\boldsymbol{\mathcal{L}oss} = \begin{cases} \boldsymbol{\mathcal{L}}_{task} + 0.01\boldsymbol{\mathcal{L}}_{load}, if \frac{\boldsymbol{\mathcal{L}}_{load}}{\boldsymbol{\mathcal{L}}_{task}} > 0.5 \\ \boldsymbol{\mathcal{L}}_{task} + 0.1\boldsymbol{\mathcal{L}}_{load}, else \end{cases} \quad (12)$$

This adaptive weighting strategy stabilizes training and ensures a better balance between classification accuracy and expert diversity, ultimately contributing to improved model generalization and robustness.

Models	F1	Acc	Pre	Spec
LateFusion (Yu et al. 2019)	0.869	0.792	0.829	0.927
Loose (Wang et al. 2019)	0.897	0.837	–	–
CVSA (Lin et al. 2025a)	0.850	0.762	0.793	0.917
Yoo (Yoo et al. 2019)	0.792	0.690	–	–
MM-CNN (Wang et al. 2022)	0.872	0.804	0.869	0.933
MM-CNN_da (Wang et al. 2022)	<u>0.914</u>	<u>0.863</u>	<u>0.900</u>	<u>0.950</u>
smartDSP (Wu et al. 2023)	0.894	0.832	0.873	0.942
MVCINN (Luo et al. 2023)	0.737	0.611	0.669	0.867
MVCNN (Su et al. 2015)	0.858	0.776	0.810	0.921
ClipResnet (Radford et al. 2021)	0.906	0.853	0.895	0.949
EYEMOST (Zou et al. 2024)	0.909	0.853	0.882	0.948
Ours	**0.925**	**0.881**	**0.911**	**0.959**

Table 1: Comparison of state-of-art models on the MMAD dataset.

Experiments

Experimental Setups

Datasets We evaluate our method on two publicly available multimodal retinal datasets. The first dataset is the Multimodal AMD Dataset (MMAD), collected by the Department of Ophthalmology at Peking Union Medical College Hospital and made publicly available by (Wang et al. 2022). MMAD contains expert-annotated 1,094 CFP and 1,289 OCT images, categorized into four classes: normal,

Model	normal			dryAMD			PCV			wetAMD		
	Sen	Spe	F1	Sen	Spe	F1	Sen	Spe	F1	Sen	Spe	F1
LateFusion (Yu et al. 2019)	**1.000**	**1.000**	**1.000**	**1.000**	0.943	**0.971**	0.553	0.948	0.698	0.790	0.827	0.808
CVSA (Lin et al. 2025a)	**1.000**	0.991	0.995	0.789	0.942	0.859	0.638	0.906	0.749	0.763	0.828	0.794
MM-CNN (Wang et al. 2022)	**1.000**	**1.000**	**1.000**	0.737	0.997	0.847	0.787	0.910	0.841	0.790	0.819	0.801
MM-CNN_da (Wang et al. 2022)	**1.000**	**1.000**	**1.000**	0.868	**1.000**	0.929	0.794	0.948	0.864	0.868	0.860	0.864
smartDSP (Wu et al. 2023)	**1.000**	**1.000**	**1.000**	0.921	0.990	0.954	0.659	0.958	0.781	0.868	0.819	0.843
MVCINN (Luo et al. 2023)	0.900	**1.000**	0.947	0.750	0.757	0.754	0.558	0.886	0.685	0.425	0.824	0.560
MVCNN (Su et al. 2015)	**1.000**	**1.000**	**1.000**	0.947	0.980	0.963	0.659	0.875	0.752	0.631	0.828	0.716
ClipResnet (Radford et al. 2021)	**1.000**	**1.000**	**1.000**	0.815	**1.000**	0.898	0.766	**0.968**	0.855	**0.921**	0.828	0.872
EYEMOST+T (Zou et al. 2024)	**1.000**	**1.000**	**1.000**	0.894	**1.000**	0.944	0.787	0.927	0.851	0.815	**0.867**	0.840
Ours	**1.000**	**1.000**	**1.000**	0.842	**1.000**	0.914	**0.829**	**0.968**	**0.893**	**0.921**	**0.867**	**0.893**

Table 2: Comparison of the models across the diseases on the MMAD dataset.

Models	**Acc**	**Kappa**
B-EF (Hua et al. 2020)	0.660	0.456
M^2LC (Woo et al. 2018)	0.710	0.527
EyeStar (Wu et al. 2023)	0.860	0.774
MCDO (Kendall and Gal 2017)	0.758	0.636
DE (Lakshminarayanan, Pritzel, and Blundell 2017)	0.710	0.539
TMC (Han et al. 2022)	0.810	0.658
CVSA (Lin et al. 2025a)	0.670	0.606
MM-CNN (Wang et al. 2022)	0.840	0.810
SmartDSP (Wu et al. 2023)	0.840	0.743
MVCINN (Luo et al. 2023)	0.640	0.456
MVCNN (Su et al. 2015)	0.840	0.797
ClipResnet (Radford et al. 2021)	0.810	0.758
EyeMoSt+T (Zou et al. 2024)	0.820	0.732
EyeMoSt+C (Zou et al. 2024)	0.860	0.761
Ours	**0.870**	**0.841**

Table 3: Comparison of state-of-art models on the GAMMA dataset.

dry AMD, polypoidal choroidal vasculopathy (PCV), and wet AMD. For consistency and fair comparison, we adopt the same training, validating and testing split protocol as described in (Wang et al. 2022).

The second dataset is the GAMMA dataset for glaucoma recognition, consisting of 100 paired CFP-OCT cases, each labeled with a three-level glaucoma severity grade. Further details regarding dataset collection and labeling are provided in (Wu et al. 2023). Following the protocol in (Zou et al. 2024), we split the dataset into 80% training and 20% testing cases. To ensure robust and unbiased performance evaluation, we implement a rigorous five-fold cross-validation strategy throughout all experiments.

Model details We adopt ResNet-18 as the backbone for feature extraction in both modalities, initialized with weights pre-trained on ImageNet. To accommodate the modality-specific input characteristics, we modify the first convolutional block of the OCT ResNet-18 branch for the GAMMA dataset, where OCT inputs consist of 256 slices, in contrast to the 3-channel OCT inputs used in MMAD. In the CMER module, the total number of CNN experts for each modality is set to $N = 6$, with the top $K = 2$ experts selected dynamically based on cross-modal routing signals. For wavelet decomposition and reconstruction, the Haar wavelet is used as the basis function due to its simplicity and effectiveness in capturing local frequency patterns. The cross-attention mechanism employed in the TEWF module is configured with 8 attention heads, a depth of 4 layers, and a token embedding dimension of 1024.

Evaluation metrics To comprehensively assess the performance of multimodal fundus disease classification, we adopt several widely recognized evaluation metrics, including accuracy (Acc), F1 score (F1), Cohen's kappa score (Ka), precision (Pre), sensitivity (Sen), and specificity (Spe).

Compared methods We evaluate our method against a comprehensive set of state-of-the-art (SOTA) approaches on both multimodal retinal datasets. For the MMAD dataset, we compare with MM-CNN (Wang et al. 2022), the current SOTA method, along with its variants including MM-CNN_Loose (Wang et al. 2019) and MM-CNN_da. On the GAMMA dataset, we benchmark against EYEMOST (Zou et al. 2024), the SOTA method in the GAMMA dataset, and its variants EYEMOST+T and EYEMOST+C. We also include the top two competition-winning methods in GAMMA competition, Eyestar and SmartDSP (Wu et al. 2023). To ensure a thorough evaluation, we include additional multimodal fusion baselines such as LateFusion (Yu et al. 2019), Yoo (Yoo et al. 2019), B-EF (Hua et al. 2020), M^2LC (Woo et al. 2018), MCDO (Kendall and Gal 2017), DE (Lakshminarayanan, Pritzel, and Blundell 2017), and TMC (Han et al. 2022). Given the relevance of transformer-based architectures in recent retinal imaging research, we also compare with SOTA feature-fusion transformer models including CVSA (Lin et al. 2025a) and MVCINN (Luo et al. 2023), as well as the classic multi-view feature fusion model MVCNN (Su et al. 2015), which has been widely adopted in natural image processing. In addition, we evaluate a vision-language pretraining baseline using CLIP (Radford et al. 2021), where the ResNet backbone is trained with CLIP-based alignment loss, referred to as ClipResNet, to assess the potential of contrastive multimodal supervision.

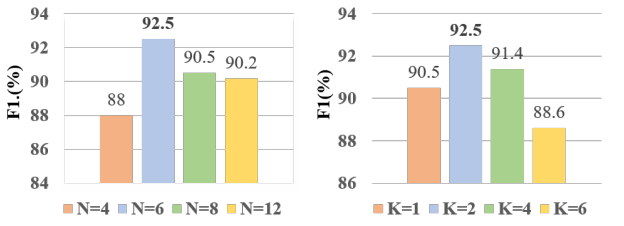
Main results

Results on MMAD dataset We report quantitative comparisons on the MMAD dataset in Table 1, benchmarking our method against a wide range of SOTA multimodal models. Our proposed method outperforms all baselines across all evaluation metrics, achieving the highest F1 score (0.925), accuracy (0.881), precision (0.911), and specificity (0.959). Compared to the best-performing prior method MM-CNN_da, our model shows a +1.1% improvement in F1 and a +1.8% improvement in accuracy, while further increasing specificity by 0.9%. Notably, even compared to the strong CLIP-based backbone (ClipResnet), which benefits from cross-modal aligning, our framework demonstrates superior performance, highlighting the effectiveness of the proposed CMER and TEWF modules in capturing fine-grained, complementary features across modalities.

To further assess class-wise performance, Table 2 presents a detailed comparison of sensitivity, specificity, and F1-score across the four disease categories in the MMAD dataset: normal, dry AMD, PCV, and wet AMD. All compared methods achieve perfect scores for the normal class, reflecting the relative ease of distinguishing healthy cases. However, notable performance disparities emerge in the more challenging pathological categories. Our method still achieves the highest F1-score, sensitivity and specificity on both the PCV and the wet AMD class, outperforming all baselines. For dry AMD, our approach also achieves the highest specificity (100%). These results highlight the effectiveness of our adaptive fusion strategy in capturing lesion-specific characteristics and enhancing class discrimination, particularly for difficult-to-diagnose retinal diseases where fine-grained pathological cues are essential for accurate classification.

Models	F1	Acc	Pre	Spec
w/o CMER	0.901	0.846	0.892	0.946
w/o WT	0.914	0.874	0.886	0.950
w/o CA	0.907	0.853	0.887	0.949
w/o CMER & TEWF	0.894	0.832	0.873	0.942
Ours	**0.925**	**0.881**	**0.911**	**0.959**

Table 4: Ablation studies in the proposed method.



(a) Total number of experts (K=2) (b) Number of routed experts (N=6)

Figure 3: Hyperparameter analysis

Results on GAMMA dataset Table 3 presents a comparative analysis of our method against a broad range of SOTA approaches on the GAMMA dataset. Our method achieves the highest overall accuracy of 0.870 and the highest Cohen's Kappa score of 0.841, surpassing all competing methods. Compared to the best-performing baseline in terms of Kappa (MM-CNN, 0.810), our model improves agreement by +3.1%. While EyeMoSt+C match our accuracy (0.860), it fall behind in Kappa, indicating lower inter-class consistency and reliability. The consistently strong performance of our model across both metrics highlights the effectiveness of our cross-modal expert routing and wavelet-guided fusion mechanisms in capturing subtle, clinically relevant patterns in glaucoma grading.

Ablation Studies and Hyperparameter Evaluations Ablations in Table 4 show each module is essential. Removing CMER ("w/o CMER") markedly degrades all metrics, validating its adaptive modality-specific routing. Skipping WT ("w/o WT") and applying cross-attention directly to Top-K features reduces accuracy, indicating the value of multi-scale wavelet decomposition. Dropping cross-attention and merely summing modalities also harms performance ("w/o CA"), confirming the need for dynamic interaction. And the baseline model without both CMER and TEWF ("w/o CMER & TEWF") yields the lowest performance, underscoring the complementary strengths of these two components.

We further analyze the impact of the total number of experts (N) and the number of routed experts (K) on model performance. As shown in Figure 3, the configuration with $N = 6$ achieves the highest F1 score of 92.5%, indicating an optimal balance between model capacity and feature learning. When $N = 6$, selecting $K = 2$ yields the best performance among all tested values, suggesting that routing to two experts effectively facilitates multimodal information fusion while avoiding redundancy.

Conclusions

In this paper, we propose a novel multimodal fusion framework for retinal disease diagnosis, which dynamically leverages modality-specific contributions through CMER and TEWF. Our approach effectively addresses the limitations of existing fusion methods by adaptively modulating cross-modal interactions based on contextual and frequency-specific information. The integration of CNN-based experts, guided by contextual cues from complementary modalities, ensures that only the most relevant features contribute to the fusion process. Moreover, the use of wavelet decomposition and high-frequency attention enhances the sensitivity to fine-grained pathological details, crucial for accurate disease recognition. Experimental evaluations on two publicly available multimodal retinal disease datasets, MMAD and GAMMA, demonstrate the superiority of our method over existing state-of-the-art approaches, achieving significant improvements in classification accuracy, sensitivity, and robustness. Our method not only sets a new benchmark for retinal disease recognition but also provides a framework that can be adapted to other medical tasks where complementary information from different modalities is essential.

Acknowledgments

This work was supported by Shenzhen Medical Research Fund C2501016, National Natural Science Foundation of China under Grant No. 62502320, Science and Technology Innovation Committee of Shenzhen Municipality under grant no. JCYJ20250604145426036, National Natural Scientific Foundation of China under Grant 62102339 and the Guangdong Major Project of Basic and Applied Basic Research under Grant 2023B0303000010.

References

Bilal, H.; Keles, A.; and Bendechache, M. 2025. Advances in disease detection through retinal imaging: A systematic review. *Computers in Biology and Medicine*, 194: 110412.

Cao, B.; Sun, Y.; Zhu, P.; and Hu, Q. 2023. Multi-modal gated mixture of local-to-global experts for dynamic image fusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, 23555–23564.

Cao, L.; Li, H.; Zhang, Y.; Zhang, L.; and Xu, L. 2020. Hierarchical method for cataract grading based on retinal images using improved Haar wavelet. *Information Fusion*, 53: 196–208.

Ferris, F. L.; and Tielsch, J. M. 2004. Blindness and visual impairment: a public health issue for the future as well as today. *Archives of Ophthalmology*, 122(4): 451–452.

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.

Hua, C.-H.; Kim, K.; Huynh-The, T.; You, J. I.; Yu, S.-Y.; Le-Tien, T.; Bae, S.-H.; and Lee, S. 2020. Convolutional network with twofold feature augmentation for diabetic retinopathy recognition from multi-modal images. *IEEE Journal of Biomedical and Health Informatics*, 25(7): 2686–2697.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Lei, X.; Chen, Z.; Liu, H.; Chen, J.; Tan, H.; Dai, W.; Wang, X.; and Xu, H. 2024. A Cross-Modal Feature Fusion Method to Diagnose Macular Fibrosis in Neovascular Age-Related Macular Degeneration. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, 1–5. IEEE.

Li, S.; Zhang, D.; Li, X.; Ou, C.; An, L.; Xu, Y.; Yang, W.; Zhang, Y.; and Cheng, K.-T. 2024. Vessel-promoted OCT to OCTA image translation by heuristic contextual constraints. *Medical Image Analysis*, 98: 103311.

Lin, Y.; Dou, X.; Luo, X.; Wu, Z.; Liu, C.; Luo, T.; Wen, J.; Ling, B. W.-k.; Xu, Y.; and Wang, W. 2025a. Multi-view diabetic retinopathy grading via cross-view spatial alignment and adaptive vessel reinforcing. *Pattern Recognition*, 164: 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.

Lin, Z.; He, Z.; Wang, X.; Su, W.; Tan, J.; Deng, Y.; and Xie, S. 2025c. Cross-scale fuzzy holistic attention network for diabetic retinopathy grading from fundus images. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

Liu, W.; Zhang, Y.; Wang, X.; and Zhang, L. 2025. Deep Multi-Level Contrastive Clustering for Multi-Modal Remote Sensing Images. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 1239–1247.

Luo, X.; Chen, P.; Liu, C.; Jin, X.; Wen, J.; Liu, Y.; and Wang, J. 2025a. Enhancing Multimodal Protein Function Prediction Through Dual-Branch Dynamic Selection with Reconstructive Pre-Training. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, 7598–7606.

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.; Wang, W.; Xu, Y.; Lai, Z.; Jin, X.; Zhang, B.; and Zhang, D. 2024. A deep convolutional neural network for diabetic retinopathy detection via mining local and long-range dependence. *CAAI Transactions on Intelligence Technology*, 9(1): 153–166.

Luo, X.; Xu, Q.; Wu, H.; Liu, C.; Lai, Z.; and Shen, L. 2025b. Like an Ophthalmologist: Dynamic Selection Driven Multi-View Learning for Diabetic Retinopathy Grading. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 19224–19232.

Qin, Y.; Feng, G.; and Zhang, X. 2025. Scalable One-Pass Incomplete Multi-View Clustering by Aligning Anchors. In *Advancement of Artificial Intelligence*, 20042–20050.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

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.

Su, W.; Tang, S.; Liu, X.; Yi, X.; Ye, M.; Zu, C.; Li, J.; and Zhu, X. 2025. Domain Adaptive Diabetic Retinopathy Grading with Model Absence and Flowing Data. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 28337–28346.

Wang, L.; Qi, C.; Ou, C.; An, L.; Jin, M.; Kong, X.; and Li, X. 2024. MultiEYE: Dataset and Benchmark for OCT-Enhanced Retinal Disease Recognition from Fundus Images. *IEEE Transactions on Medical Imaging*.

Wang, W.; Li, X.; Xu, Z.; Yu, W.; Zhao, J.; Ding, D.; and Chen, Y. 2022. Learning two-stream CNN for multi-modal age-related macular degeneration categorization. *IEEE Journal of Biomedical and Health Informatics*, 26(8): 4111–4122.

Wang, W.; Xu, Z.; Yu, W.; Zhao, J.; Yang, J.; He, F.; Yang, Z.; Chen, D.; Ding, D.; Chen, Y.; and Li, X. 2019. Two-Stream CNN with Loose Pair Training for Multi-modal AMD Categorization. In *MICCAI*, 156–164.

Wen, C.; Ye, M.; Li, H.; Chen, T.; and Xiao, X. 2024. Concept-based Lesion Aware Transformer for Interpretable Retinal Disease Diagnosis. *IEEE Transactions on Medical Imaging*.

Wen, Y.; Luo, B.; Shi, W.; Ji, J.; Cao, W.; Yang, X.; and Sheng, B. 2025. Sat-net: structure-aware transformer-based attention fusion network for low-quality retinal fundus images enhancement. *IEEE Transactions on Multimedia*.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Wu, J.; Fang, H.; Li, F.; Fu, H.; Lin, F.; Li, J.; Huang, Y.; Yu, Q.; Song, S.; Xu, X.; et al. 2023. Gamma challenge: glaucoma grading from multi-modality images. *Medical Image Analysis*, 90: 102938.

Xie, L.; Luan, T.; Cai, W.; Yan, G.; Chen, Z.; Xi, N.; Fang, Y.; Shen, Q.; Wu, Z.; and Yuan, J. 2025. dFLMoE: Decentralized Federated Learning via Mixture of Experts for Medical Data Analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 10203–10213.

Yang, J.; Mehta, N.; Demirci, G.; Hu, X.; Ramakrishnan, M. S.; Naguib, M.; Chen, C.; and Tsai, C.-L. 2024. Anomaly-guided weakly supervised lesion segmentation on retinal OCT images. *Medical Image Analysis*, 94: 103139.

Yoo, T. K.; Choi, J. Y.; Seo, J. G.; Ramasubramanian, B.; Selvaperumal, S.; and Kim, D. W. 2019. The possibility of the combination of OCT and fundus images for improving the diagnostic accuracy of deep learning for age-related macular degeneration: a preliminary experiment. *Medical & biological engineering & computing*, 57: 677–687.

Yu, J.; Li, J.; Yu, Z.; and Huang, Q. 2019. Multimodal transformer with multi-view visual representation for image captioning. *IEEE transactions on circuits and systems for video technology*, 30(12): 4467–4480.

Zhang, P.; Li, Y.; Song, H.; Jiang, Y.; Tao, Y.; Lin, H.; and Cui, H. 2025. SIGraph: Saliency Image-Graph Network for Retinal Disease Classification in Fundus Image. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 10049–10057.

Zhang, X.; Xiao, Z.; Yang, B.; Wu, X.; Higashita, R.; and Liu, J. 2024a. Regional context-based recalibration network for cataract recognition in AS-OCT. *Pattern Recognition*, 147: 110069.

Zhang, Y.; Yan, S.; Jiang, X.; Zhang, L.; Cai, Z.; and Li, J. 2024b. Dual graph learning affinity propagation for multimodal remote sensing image clustering. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–13.

Zhang, Y.; Yan, S.; Zhang, L.; and Du, B. 2024c. Fast projected fuzzy clustering with anchor guidance for multimodal remote sensing imagery. *IEEE Transactions on Image Processing*.

Zou, K.; Lin, T.; Han, Z.; Wang, M.; Yuan, X.; Chen, H.; Zhang, C.; Shen, X.; and Fu, H. 2024. Confidence-aware multi-modality learning for eye disease screening. *Medical Image Analysis*, 96: 103214.