Addressing Gradient Conflicts in Multimodal Fundus Disease Recognition with Fusion-Guided Learning

Xiaozhou Liu[1*], Zhike Qiu[1*], Luping Zeng[1], Jiahui Hu[1], Jiahao Liu[1], Liangming Wen[1†], and Zhiguo Du[1]

Business College, Southwest University, Chongqing, China
wlm20240211@swu.edu.cn

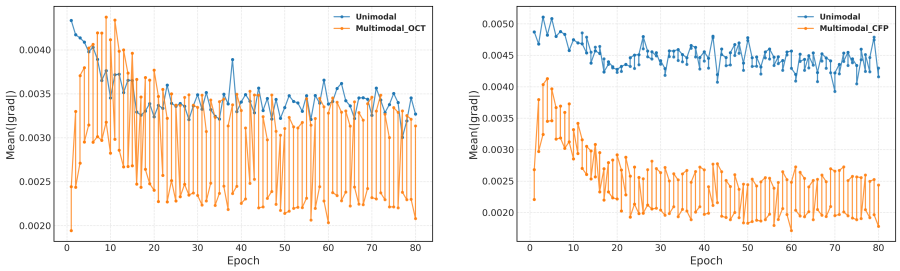
Abstract. Multi-modal fundus image analysis holds significant clinical value for the early diagnosis of ophthalmic diseases. While existing studies on multi-modal diagnosis predominantly focus on designing complex fusion architectures, they often overlook the underlying optimization dynamics during joint training. In this work, we reveal that optimization conflicts between the OCT and CFP encoders and the fusion module lead to suboptimal model performance. To address this, we propose a novel framework incorporating Fusion-Guided Gradient Decoupling Learning (FGDL) and a Cross-Modal Semantic Alignment (CMSA) module. By decoupling optimization paths and employing conflict-aware projection to rectify uni-modal gradients, FGDL eliminates inter-modal gradient interference while guiding uni-modal features toward the global multi-modal optimum. Meanwhile, the CMSA module achieves semantic alignment between OCT structural features and CFP texture features by capturing global dependencies. Furthermore, we introduce a loose pairing strategy to enhance the model's robustness against data heterogeneity. Experiments on the MMC-AMD and GAMMA datasets demonstrate that our method achieves accuracies of 93.01% and 95.00%, respectively, significantly outperforming existing state-of-the-art approaches. Code is available at https://github.com/lxz8763/FGDL-Fundus.

Keywords: Multimodal learning · Fundus disease recognition · Gradient decoupling

1 Introduction

Fundus diseases are a leading cause of irreversible blindness worldwide. Therefore, timely and accurate diagnosis is crucial for early intervention and treatment [13,17]. Color Fundus Photography (CFP) provides texture and vascular information of the retinal surface, while Optical Coherence Tomography (OCT) characterizes deep structural changes [7,12]. The two modalities exhibit significant complementarity in pathological representations [14,5,2].

[*] Equal contribution.
[†] Corresponding author.



(a) OCT encoder (b) CFP encoder

Fig. 1. Comparative analysis of mean gradients for the OCT and CFP branches under unimodal and multimodal training.

Inspired by clinical diagnostic workflows, the computer-aided diagnosis community has recently witnessed a surge in multimodal deep learning frameworks tailored for fundus diseases [24,22,18,28,25,9]. Current mainstream methods typically adopt a two-stream network architecture, utilizing two independent networks to extract feature representations from OCT and CFP respectively, followed by modality fusion via feature concatenation, weighted summation, or attention mechanisms [19,24,8]. To further enhance the efficiency of feature interaction, Yu et al. proposed the MSMAE algorithm, which utilizes keypoint detection and supervoxel segmentation to extract modality-specific multi-perceptual features and introduces a Graph Transformer for structural modeling [27]. Zheng et al. incorporated evidential deep learning to quantify predictive uncertainty, achieving reliable multi-view fusion [29].

However, most existing studies focus on designing increasingly complex fusion frameworks while neglecting the underlying optimization dynamics during joint multimodal training [22,29,28,27,15]. Traditional joint optimization strategies tightly couple the updates of feature extractors and fusion modules. This strong coupling can induce optimization conflicts during backpropagation, where the fusion module suppresses the gradients backpropagated to the modality encoders. Consequently, unimodal encoders are inadequately updated, preventing the multimodal model from fully realizing its performance potential [23]. As illustrated in Fig. 1, during joint multimodal training, the magnitude of gradients backpropagated to the CFP encoder is significantly lower than that of the unimodal training baseline, while the gradients of the OCT encoder exhibit severe oscillation and instability. This confirms that traditional joint optimization strategies lead to inter-modality optimization conflicts and gradient interference.

Beyond optimization issues, limitations in feature representation pose another critical challenge restricting model performance. Fundus disease lesions often exhibit global correlations; relying solely on local convolutions makes it difficult to effectively model such cross-modal global semantic co-occurrences [14,21,16]. Furthermore, in real-world clinical scenarios, acquiring strictly paired multi-

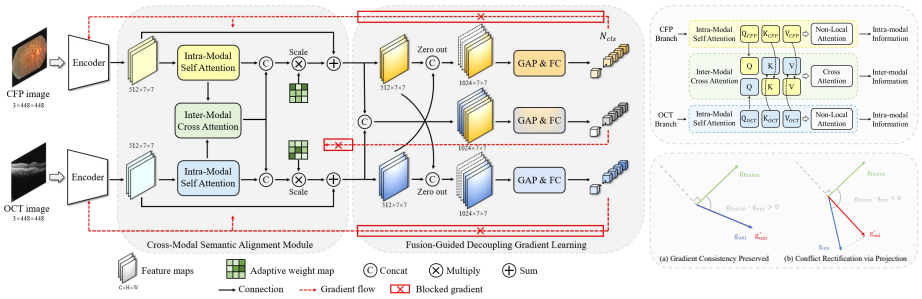


Fig. 2. The overall architecture of the proposed multimodal fundus disease recognition network. The left panel illustrates the backbone network alongside the two core modules, FGDL and CMSA. The right panel displays the detailed structure of the CMSA (Right-Top) and a geometric illustration of the projection mechanism for gradient conflict correction in FGDL (Right-Bottom).

modal data is exceedingly difficult and expensive. To improve the effectiveness of model training on limited datasets, more flexible data augmentation strategies must be introduced [1,18].

To address these challenges, this paper proposes a two-stream fundus disease recognition network based on fusion-guided gradient decoupling learning. Specifically, to tackle the optimization conflict problem, we propose a novel collaborative optimization mechanism, fusion-guided gradient decoupling learning (FGDL). Additionally, we integrate cross-modal semantic alignment (CMSA) into the dual-branch ResNet architecture.

The main contributions of this paper are summarized as follows: 1) We propose a novel multimodal learning framework for fundus imaging, termed FGDL. By combining gradient decoupling with a conflict-aware projection mechanism, this framework effectively resolves gradient competition and suppression during the fusion of OCT and CFP, while ensuring consistency between unimodal and multimodal optimization directions; 2) We design a CMSA module. By capturing global long-range dependencies, it effectively uncovers latent correlations between lesion regions across different modalities, significantly facilitating the semantic alignment of cross-modal features; 3) Combined with a loose pairing data augmentation strategy, we conduct extensive experiments on two public datasets, MMC-AMD and GAMMA. Our method outperforms existing state-of-the-art (SOTA) models in accuracy and other key evaluation metrics.

2 Methodology

Let the multimodal fundus disease dataset be denoted as $\mathcal{D} = \{(x_i^{oct}, x_i^{cfp}, y_i)\}_{i=1}^{N}$, where $x_i^{oct} \in \mathbb{R}^{H\times W\times 3}$ represents an OCT image, $x_i^{cfp} \in \mathbb{R}^{H\times W\times 3}$ represents the corresponding CFP image, and $y_i \in \{1, \ldots, K\}$ denotes the disease category

label. N is the total number of samples, and K is the number of categories. Our objective is to learn a multimodal classification model $f(x^{oct}, x^{cfp}; \Theta)$ to predict the category label $\hat{y}$ for each paired sample.

2.1 Loose Pairing Strategy

During the training process, we adopt a category-consistent loose pairing strategy [19]. For a given OCT image, instead of strictly limiting its pairing to the CFP image from the same eye, we allow it to be paired with any CFP image that shares the same disease category label. This approach increases the number and diversity of valid training pairs.

2.2 Cross-Modal Semantic Alignment

To achieve semantic-level feature alignment between OCT and CFP, we propose the CMSA module. This module consists of two components: intra-modal global modeling and cross-modal semantic interaction. In both the OCT and CFP branches, we introduce a non-local self-attention mechanism to aggregate global semantic information within a single modality [20]. Let $X_{\text{oct}}, X_{\text{cfp}} \in \mathbb{R}^{C\times H\times W}$ denote the feature maps output by the backbone networks, where C, H, and W represent the number of channels, height, and width, respectively. For any modality $m \in \{\text{oct}, \text{cfp}\}$, its intra-modal enhanced feature is expressed as:

$$Y_m^{\text{self}} = \text{Softmax}\left(\frac{\theta(X_m)^\top \phi(X_m)}{\sqrt{d_k}}\right) g(X_m), \tag{1}$$

where $\theta(\cdot)$, $\phi(\cdot)$, and $g(\cdot)$ are 1×1 convolutional mapping functions that generate the Query, Key, and Value, respectively; d_k is the channel dimension of the Key. This operation performs weighted aggregation based on the semantic similarity between features, thereby modeling global dependencies within the modality.

CMSA further introduces a cross-modal attention mechanism. Using the OCT features as queries and the CFP features as keys and values, it semantically enhances the OCT representation:

$$Y_{\text{oct}}^{\text{cross}} = \text{Softmax}\left(\frac{Q_{\text{oct}} K_{\text{cfp}}^\top}{\sqrt{d_k}}\right) V_{\text{cfp}}, \tag{2}$$

where Q_{oct} is generated from the OCT features, and K_{cfp} and V_{cfp} denote the Key and Value corresponding to the CFP features, respectively. Similarly, the cross-modal enhanced feature for the CFP branch, $Y_{\text{cfp}}^{\text{cross}}$, can be obtained.

Finally, we concatenate the intra-modal global features with the cross-modal aligned features, and fuse them back into the original feature stream via a gated residual mechanism:

$$Z_m = X_m + \sigma(\mathbf{w}_{\text{gate}}) \cdot \mathcal{F}_{\text{fusion}}\left([Y_m^{\text{self}}, Y_m^{\text{cross}}]\right), \tag{3}$$

where $[\cdot,\cdot]$ denotes concatenation along the channel dimension, $\mathcal{F}_{\text{fusion}}$ is a fusion convolutional layer, and $\mathbf{w}_{\text{gate}}$ represents a learnable scalar weight.

2.3 Fusion-Guided Gradient Decoupling Learning

To alleviate the issues of gradient suppression and directional conflict during joint multimodal training, we design a FGDL strategy.

FGDL first constructs two independent optimization paths using a stop-gradient operation. When calculating the multimodal fusion loss L_{fusion}, its gradient backpropagation to the unimodal backbones is blocked, ensuring that this loss is solely used to update the fusion module and the classification head:

$$L_{\text{fusion}} = \text{CrossEntropy}(\text{Fusion}(\text{SG}(F_{\text{oct}}), \text{SG}(F_{\text{cfp}})),\, y)\,, \tag{4}$$

where F_{oct} and F_{cfp} represent the output features of the OCT and CFP encoders respectively, y is the ground-truth label, and $\text{SG}(\cdot)$ denotes the stop-gradient operation. Meanwhile, we introduce independent unimodal classification heads and corresponding loss functions, L_{oct} and L_{cfp}, for the OCT and CFP branches, respectively. The gradients generated by these losses are strictly used to update their respective branches, thereby preventing the fusion branch from suppressing the unimodal feature learning process.

Because unimodal tasks lack cross-modal contextual constraints, their optimization directions may deviate from the global objective of multimodal fusion [26]. To address this, FGDL introduces a gradient alignment mechanism prior to parameter updates to coordinate the optimization directions of both. The gradient derived from the backpropagation of the multimodal fusion loss is denoted as $\mathbf{g}_{\text{fusion}}$, serving as the global optimization reference direction. The gradient generated by the unimodal loss is denoted as $\mathbf{g}_{\text{uni}}$, representing the local optimization direction of the single modality. When their inner product $\mathbf{g}_{\text{fusion}} \cdot \mathbf{g}_{\text{uni}} < 0$ is detected, it indicates a conflict between the unimodal update direction and the fusion objective. In this case, a vector projection operation is employed to remove the component of the unimodal gradient that conflicts with the fusion gradient, yielding the corrected unimodal gradient:

$$\mathbf{g}^*_{\text{uni}} = \begin{cases} \mathbf{g}_{\text{uni}} - \frac{\mathbf{g}_{\text{uni}} \cdot \mathbf{g}_{\text{fusion}}}{\|\mathbf{g}_{\text{fusion}}\|^2}\, \mathbf{g}_{\text{fusion}}, & \mathbf{g}_{\text{fusion}} \cdot \mathbf{g}_{\text{uni}} < 0, \\ \mathbf{g}_{\text{uni}}, & \text{otherwise.} \end{cases} \tag{5}$$

This process is executed during the backpropagation phase, guiding the unimodal update direction purely at the optimization level. Ultimately, the network parameters are jointly driven by the fusion gradient and the corrected unimodal gradients. The corresponding total loss function is defined as:

$$\mathcal{L}_{\text{total}} = L_{\text{fusion}} + \alpha\,(L_{\text{oct}} + L_{\text{cfp}})\,, \tag{6}$$

where α is a balancing coefficient used to control the weight of the unimodal constraints within the overall optimization.

3 Experiments

3.1 Datasets and Experimental Setup

We conduct experiments on two publicly available and widely used multimodal datasets. The MMC-AMD dataset is a clinical dataset specifically designed for

Table 1. Comparative experimental results (%) on the MMC-AMD and GAMMA datasets.

Method	MMC-AMD				GAMMA			
	Acc	Recall	F1	Kappa	Acc	Recall	F1	Kappa
CFP-ResNet18 [3]	72.55	58.45	67.91	57.56	85.00	84.26	83.37	76.83
OCT-ResNet18 [3]	80.39	83.14	80.35	72.05	60.00	61.11	64.19	42.45
MM-CNN [19]	80.39	76.70	79.03	71.88	80.00	80.56	79.55	69.70
MSAN [4]	70.35	63.94	67.33	52.34	65.00	52.38	49.36	39.13
DeepGuide [11]	70.59	54.35	55.11	55.11	85.00	84.26	83.37	76.83
CRD-Net [10]	85.62	87.75	84.66	79.57	90.00	92.59	90.70	84.85
BCA-Trans [6]	84.30	85.80	85.98	78.68	-	-	-	-
MSAS-ViT [27]	-	-	-	-	94.00	88.50	88.00	-
Ours	**93.01**	**94.18**	**94.08**	**90.45**	**95.00**	**93.34**	**93.34**	**92.00**

fine-grained age-related macular degeneration. The data splitting follows the official settings [19]. The GAMMA Challenge dataset is a multimodal dataset for glaucoma grading [24], and we follow the conventional settings [10]. Due to its relatively small size, we randomly split this dataset twice and report the average results.

All experiments were implemented in PyTorch on an NVIDIA RTX 4090D GPU. The baseline model employed a two-stream ResNet-18 with simple feature concatenation, and input images were resized to $448 \times 448 \times 3$. The optimizer was SGD with momentum 0.9 and weight decay 1×10^{-4}. The initial learning rate was set to 1×10^{-3} and decayed using a cosine annealing schedule. Training was conducted for up to 80 epochs with an early stopping patience of 20 epochs. The hyperparameter α was set to 5.0 for MMC-AMD and 1.0 for GAMMA. Image preprocessing and data augmentation followed the protocol in [19].

3.2 Comparative Experiments

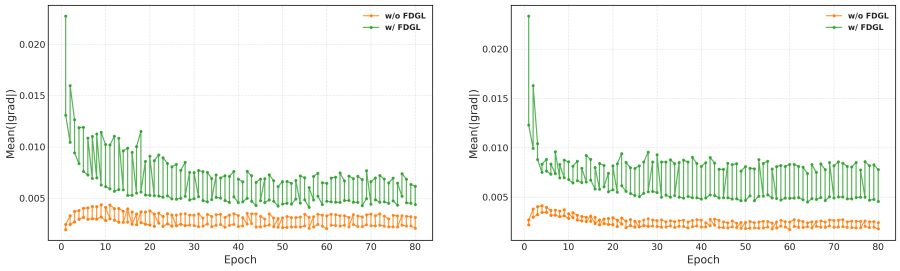
We systematically compare our framework with unimodal baseline methods and the latest SOTA multimodal methods. As shown in Table 1, our method outperforms all competing approaches on both datasets. On the MMC-AMD dataset, our model achieves a classification accuracy of 93.01% and a Kappa coefficient of 90.45%. Similarly, on the GAMMA dataset, our method yields an accuracy of 95.00% and a Kappa coefficient of 92.00%. The overall performance surpasses that of current advanced models.

3.3 Ablation Study

As shown in Table 2, we perform a step-by-step ablation study on the MMC-AMD dataset. The introduction of loose pairing significantly improves classification accuracy, which is attributed to the effective expansion of the training

Table 2. Ablation study of Loose Pairing, CMSA, and FGDL.

Loose Pairing	CMSA	FGDL	Performance (%)			
			Acc	Recall	F1	Kappa
			75.52	79.49	80.97	66.77
✓			81.82	84.91	86.43	75.34
✓	✓		83.92	86.51	88.05	78.13
✓		✓	86.71	88.63	89.74	81.86
✓	✓	✓	**93.01**	**94.18**	**94.08**	**90.45**



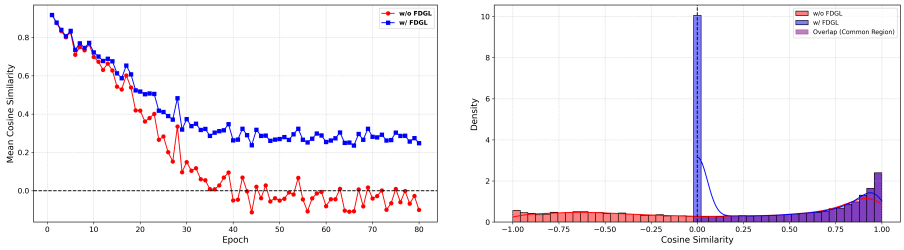
(a) OCT encoder (b) CFP encoder

Fig. 3. Evolution of the average gradient magnitude in the last layer of different modality encoders.

data scale and diversity. Employing CMSA and FGDL individually yields performance improvements of 2.1% and 4.89%, respectively. This result reveals an important phenomenon: in multimodal fundus image analysis tasks, bottlenecks at the optimization dynamics level are often more critical than the feature extraction architecture itself. By alleviating multimodal gradient conflicts, FGDL effectively unlocks the latent learning capacity of the encoders.

3.4 In-Depth Analysis of FGDL

Analysis of FGDL in Alleviating Gradient Suppression for Modality Encoders During the training process, we record the mean absolute gradient values at the last layer of both the OCT and CFP modality encoders. As illustrated in Fig. 3, without FGDL, the overall gradient magnitudes of the modality encoders in joint multimodal learning remain relatively small. After introducing FGDL, the gradient magnitudes of both modality encoders increase significantly throughout the entire training process. These results indicate that by guiding and decoupling the fusion gradients, FGDL effectively mitigates the suppressive effect of the fusion branch on the modality encoders' gradient updates, enabling the unimodal branches to maintain sufficient parameter update capacity within the multimodal framework.



(a) Temporal evolution (b) Distribution statistics

Fig. 4. Analysis of gradient direction consistency between specific modality branches and the fusion branch.

Analysis of Gradient Conflicts Between Unimodal and Fusion Branches In each training epoch, we extract the gradients backpropagated from the unimodal branch to the modality encoder, as well as those from the fusion branch to the same encoder, and calculate the cosine similarity between them to measure the consistency of gradient directions. Fig. 4(a) illustrates the average trend of the gradient cosine similarity during training. Under the baseline setting, this similarity gradually decreases as training progresses, dropping below 0 in the middle and late stages. This indicates that the optimization direction of the unimodal branch progressively deviates from the fusion objective, leading to a continuous accumulation of gradient conflicts. After incorporating FGDL, the gradient cosine similarity remains strictly positive and stabilizes at a high level throughout the entire training process. Fig. 4(b) presents the distribution statistics of the gradient cosine similarity across all batches. Under the FGDL setting, the overall gradient distribution concentrates in the positive region. This demonstrates that FGDL can continuously guide the unimodal gradients to maintain directional consistency with the multimodal fusion gradients throughout the training dynamics.

4 Conclusion

This paper presents a robust multimodal framework for fundus disease recognition, designed to address current limitations in optimization dynamics and feature representation. The proposed FGDL strategy effectively resolves gradient conflicts by guiding the unimodal encoders to optimize toward the global fusion objective without inducing suppression. Furthermore, the CMSA module enhances model performance by capturing long-range cross-modal semantic associations. Combined with the data augmentation strategy, our method achieves SOTA performance on two public datasets.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Garcea, F., Serra, A., Lamberti, F., Morra, L.: Data augmentation for medical imaging: A systematic literature review. Computers in biology and medicine **152**, 106391 (2023)
2. Gu, X., Zhou, Y., Zhao, J., Zhang, H., Pan, X., Li, B., Zhang, B., Wang, Y., Xia, S., Lin, H., et al.: Diagnostic performance and generalizability of deep learning for multiple retinal diseases using bimodal imaging of fundus photography and optical coherence tomography. Frontiers in Cell and Developmental Biology **13**, 1665173 (2025)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
4. He, X., Deng, Y., Fang, L., Peng, Q.: Multi-modal retinal image classification with modality-specific attention network. IEEE transactions on medical imaging **40**(6), 1591–1602 (2021)
5. Hemelings, R., Elen, B., Schuster, A.K., Blaschko, M.B., Barbosa-Breda, J., Hujanen, P., Junglas, A., Nickels, S., White, A., Pfeiffer, N., et al.: A generalizable deep learning regression model for automated glaucoma screening from fundus images. NPJ digital medicine **6**(1), 112 (2023)
6. Li, J., Wang, Z., Chen, Y., Zhu, C., Xiong, M., Bai, H.X.: A transformer utilizing bidirectional cross-attention for multi-modal classification of age-related macular degeneration. Biomedical Signal Processing and Control **109**, 107887 (2025)
7. Li, X., Zhou, Y., Wang, J., Lin, H., Zhao, J., Ding, D., Yu, W., Chen, Y.: Multi-modal multi-instance learning for retinal disease recognition. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 2474–2482 (2021)
8. Li, Y., El Habib Daho, M., Conze, P.H., Al Hajj, H., Bonnin, S., Ren, H., Manivannan, N., Magazzeni, S., Tadayoni, R., Cochener, B., et al.: Multimodal information fusion for glaucoma and diabetic retinopathy classification. In: International Workshop on Ophthalmic Medical Image Analysis. pp. 53–62. Springer (2022)
9. Li, Y., Hou, Q., Cao, P., Ju, J., Wang, T., Wang, M., Zou, K., Tham, Y.C., Fu, H., Zaiane, O.R.: Rifnet: Bridging modalities for accurate and detailed ocular disease analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 518–528. Springer (2025)
10. Liu, Z., Hu, Y., Qiu, Z., Niu, Y., Zhou, D., Li, X., Shen, J., Jiang, H., Li, H., Liu, J.: Cross-modal attention network for retinal disease classification based on multi-modal images. Biomedical Optics Express **15**(6), 3699–3714 (2024)
11. Mallya, M., Hamarneh, G.: Deep multimodal guidance for medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 298–308. Springer (2022)
12. Morano, J., Fazekas, B., Sükei, E., Fecso, R., Emre, T., Gumpinger, M., Faustmann, G., Oghbaie, M., Schmidt-Erfurth, U., Bogunović, H.: Multimodal foundation model and benchmark for comprehensive retinal oct image analysis. NPJ Digital Medicine **8**(1), 576 (2025)
13. Qiu, Z., Qin, Y., Zeng, L., Wen, L.: Mbfsnet: Multi-scale binocular fusion semantic co-occurrence network for multi-label fundus disease diagnosis. In: International Conference on Intelligent Computing. pp. 453–464. Springer (2025)
14. Sheng, H., Du, H., Shen, X., Wang, S., Yu, X.: Multimodal retina image analysis survey: datasets, tasks and methods. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence. pp. 10650–10659 (2025)

15. Sun, Y., Li, D., Kim, S., Wang, Y.X., Wang, J., Wong, T.Y., Liao, H., Song, S.J.: Retinal thickness prediction from multi-modal fundus photography. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 585–595. Springer (2023)
16. Tian, X., Anantrasirichai, N., Nicholson, L., Achim, A.: Tagat: topology-aware graph attention network for multi-modal retinal image fusion. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 775–784. Springer (2024)
17. Wang, L., Dai, W., Jin, M., Ou, C., Li, X.: Fundus-enhanced disease-aware distillation model for retinal disease classification from oct images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 639–648. Springer (2023)
18. Wang, L., Qi, C., Ou, C., An, L., Jin, M., Kong, X., Li, X.: Multieye: Dataset and benchmark for oct-enhanced retinal disease recognition from fundus images. IEEE Transactions on Medical Imaging **44**(4), 1711–1722 (2024)
19. Wang, W., Li, X., Xu, Z., Yu, W., Zhao, J., Ding, D., Chen, Y.: Learning two-stream cnn for multi-modal age-related macular degeneration categorization. IEEE Journal of Biomedical and Health Informatics **26**(8), 4111–4122 (2022)
20. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
21. Wang, X., Cao, A., Fan, C., Tan, Z., Wang, Y.: Dual-branch network with hybrid attention for multimodal ophthalmic diagnosis. Bioengineering **12**(6), 565 (2025)
22. Wang, Y., Zhen, L., Tan, T.E., Fu, H., Feng, Y., Wang, Z., Xu, X., Goh, R.S.M., Ng, Y., Calhoun, C., et al.: Geometric correspondence-based multimodal learning for ophthalmic image analysis. IEEE Transactions on Medical Imaging **43**(5), 1945–1957 (2024)
23. Wei, S., Luo, C., Luo, Y.: Boosting multimodal learning via disentangled gradient learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22879–22888 (2025)
24. Wu, J., Fang, H., Li, F., Fu, H., Lin, F., Li, J., Huang, Y., Yu, Q., Song, S., Xu, X., et al.: Gamma challenge: glaucoma grading from multi-modality images. Medical Image Analysis **90**, 102938 (2023)
25. Yu, K., Zhou, Y., Bai, Y., Soh, Z.D., Xu, X., Goh, R.S.M., Cheng, C.Y., Liu, Y.: Urfound: Towards universal retinal foundation models via knowledge-guided masked modeling. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 753–762. Springer (2024)
26. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient surgery for multi-task learning. Advances in neural information processing systems **33**, 5824–5836 (2020)
27. Yu, Y., Zhu, H., Qian, T., Chen, N., Huang, B.: Modality-specificity multi-aware evidence fusion algorithm using cfp and oct for fundus diseases diagnosis. Pattern Recognition **169**, 111957 (2026)
28. Zhang, Y., Xu, G., Zhao, M., Wang, H., Shi, F., Chen, S.: Tdsf-net: Tensor decomposition-based subspace fusion network for multimodal medical image classification. IEEE Transactions on Neural Networks and Learning Systems (2025)
29. Zheng, X., Wang, W., Lv, Z., Li, N.: Glaucoma progression prediction using multimodal data and trusted multi-view learning. Expert Systems with Applications **288**, 128225 (2025)