JOINT SEGMENTATION AND GRADING WITH ITERATIVE OPTIMIZATION FOR MULTIMODAL GLAUCOMA DIAGNOSIS

Zhiwei Wang[1,2], Yuxing Li[1], Meilu Zhu[1], Defeng He[2], Edmund Y. Lam[1,]*

[1]Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, China
[2]College of Information Engineering, Zhejiang University of Technology, Hangzhou, China

2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI) | 979-8-3315-7763-6/26/$31.00 ©2026 IEEE | DOI: 10.1109/ISBI61048.2026.11515543

ABSTRACT

Accurate diagnosis of glaucoma is challenging, as early-stage changes are subtle and often lack clear structural or appearance cues. Most existing approaches rely on a single modality, such as fundus or optical coherence tomography (OCT), capturing only partial pathological information and often missing early disease progression. In this paper, we propose an iterative multimodal optimization model (IMO) for joint segmentation and grading. IMO integrates fundus and OCT features through a mid-level fusion strategy, enhanced by a cross-modal feature alignment (CMFA) module to reduce modality discrepancies. An iterative refinement decoder progressively optimizes the multimodal features through a denoising diffusion mechanism, enabling fine-grained segmentation of the optic disc and cup while supporting accurate glaucoma grading. Extensive experiments show that our method effectively integrates multimodal features, providing a comprehensive and clinically significant approach to glaucoma assessment. Source codes are available at https://github.com/warren-wzw/IMO.git.

***Index Terms*—** Glaucoma, Iterative optimization, multimodal data, Segmentation, Grading

1. INTRODUCTION

High ocular disease prevalence has made automated diagnosis a key research focus [1, 2]. Despite progress, accurate glaucoma diagnosis remains challenging due to subtle structural and morphological changes [3]. Therefore, developing precise and clinically meaningful assessment methods remains a vital yet difficult task.

Various single-modality segmentation [4, 5] and grading [6] methods have been proposed for glaucoma diagnosis using fundus or OCT data. For instance, Liang et al. [6] utilized a vision transformer and graph neural network for grading. Cho et al. [5] developed an unsupervised domain adaptation framework for segmentation. To overcome the inherent limitations of single-modality data, multimodal strategies combining fundus and OCT have emerged [7, 8]. Cai et al. [7] converted OCT volumes into retinal thickness maps to complement fundus images, enhancing grading. However, current multimodal methods treat segmentation and grading as independent tasks, failing to exploit their inherent correlation and the potential for joint optimization.

To more effectively leverage multimodal complementarities and harness the potential of joint multi-task optimization [9, 10], we propose IMO, a novel model for glaucoma diagnosis. The framework processes fundus and OCT data, adopting mid-level fusion with dual encoders to extract features. To address modality discrepancies, a CMFA module aligns representations in the shared feature space. An efficient grading head and iterative segmentation decoder are also introduced to progressively refine results. Through stepwise iterative optimization, IMO jointly optimizes segmentation and grading, achieving 93.33% mDice and 81.48% precision on the GAMMA dataset, demonstrating the effectiveness of multi-task optimization. The main contributions of this paper can be summarized as follows:

- We propose IMO, a joint framework integrating fundus and OCT images for concurrent glaucoma grading and optic disc/cup segmentation.
- We develop a CMFA module for cross-modality feature alignment and a diffusion-based decoder for iterative, progressive segmentation refinement.

2. PROPOSED METHOD

2.1. Overall Framework

Fig. 1 illustrates the framework. Mid-level fusion integrates fundus and OCT features via dual encoders for modality-specific representations.

A feature pyramid encoder extracts multi-scale fundus features using progressive downsampling and pointwise convolution. Simultaneously, a network captures complementary volumetric OCT cues. The CMFA module aligns these encoder outputs into a unified representation, integrating structural and appearance cues. These fused features are then fed into two parallel task-specific branches: an iterative refinement decoder for optic disc/cup segmentation and an efficient

Corresponding author: Edmund Y. Lam (elam@eee.hku.hk)

This work is supported by the Innovation and Technology Fund (ITS/341/23) of Hong Kong, China.

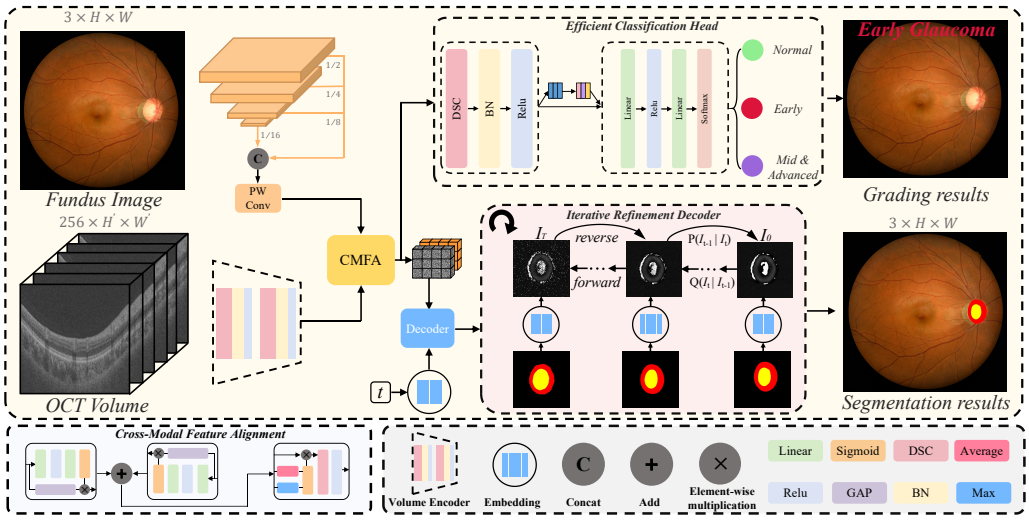

Fig. 1. The architecture of the proposed iterative multimodal optimization model.

classification head.

The classification branch predicts glaucoma stages from fused features. Depthwise separable convolutions (DSC), batch normalization, and ReLU efficiently extract spatial-channel information, followed by a SE block to model channel dependencies. Finally, a two-layer fully connected head and softmax layer output stage probabilities.

2.2. Cross-Modal Feature Alignment Module

Fundus and OCT volumes differ, leading to misaligned features. To enhance fused representations, we design CMFA to align and improve expressiveness, as shown in Fig. 1.

Specifically, CMFA first applies channel attention to each modality. Global Average Pooling (GAP) captures global channel statistics, and a two-layer fully connected network generates channel-wise attention weights W_1 and W_2 to emphasize informative channels and suppress redundant ones. This captures modality-specific differences along the channel dimension, ensuring that important features are prioritized during fusion. The operation is formulated as follows:

$$X_{\text{att}} = X \odot \sigma(W_2 \cdot ReLU(W_1 \cdot GAP(X))), \quad (1)$$

where X denotes the input feature map, σ denotes the sigmoid function and $\odot$ means element-wise multiplication.

The channel-refined features are then summed to form an initial fused representation. To further enhance spatial structural information, a spatial attention mechanism is introduced. By computing channel-wise average and max pooling, a spatial attention map $S = \sigma(Conv(Mean(X_{att}^{f+o}) \oplus Max(X_{att}^{f+o}))$ is generated to highlight the salient regions in the feature map. This step emphasizes key spatial regions across modalities and can be expressed as:

$$F_{fused}^{spatial} = X_{att}^{f+o} \odot S. \quad (2)$$

Finally, a pointwise convolution followed by ReLU activation integrates channel information and introduces nonlinearity, yielding the final fused feature. This output contains complementary information from both modalities and benefits from refined channel and spatial selection, providing a stronger representation for downstream tasks:

$$F_{\text{out}} = ReLU(Conv_{1\times1}(F_{fused}^{spatial})). \quad (3)$$

2.3. Iterative Refinement Decoder

Traditional segmentation methods often adopt a single-step segmentation paradigm, which is simple and efficient but typically struggles in complex multimodal segmentation scenarios. Inspired by the success of DDPM [11], which progressively refines predictions through multistep iterations, we designed an iterative optimization decoder for the multimodal segmentation task. Its overall architecture is illustrated in Fig. 1. In the forward process, we follow the standard noise injection scheme of DDPM:

$$X_t = \sqrt{\overline{\alpha}_t} X_0 + \sqrt{1-\overline{\alpha}_t}\epsilon, \quad (4)$$

where $\bar{\alpha}_T = \prod_{i=1}^{T} \alpha_i$, and α_i controls the intensity of the noise at each time step.

During the reverse denoising phase, we adopt the DDIM sampling strategy to accelerate the inference process. The sampling process can be expressed as:

$$x_{t-n} = \sqrt{\overline{\alpha}_{t-n}}\hat{\mathbf{x}}_0 + \sqrt{1-\overline{\alpha}_{t-n}}\epsilon_\theta(\mathbf{x}_t, t), \quad (5)$$

where $\hat{\mathbf{x}}_0$ denotes the clean segmentation representation predicted in the timestep t, $\overline{\alpha}_{t-n}$ is the cumulative product of the noise scheduling coefficients controlling the denoising trajectory, and $\epsilon_\theta(\mathbf{x}_t, t)$ represents the noise estimated by the diffusion model at step t.

Methods	Unlabel	Optic Disc	Optic Cup	mDice↑
	Dice↑	Dice↑	Dice↑	
UNet++ [13]	99.94	84.27	85.12	89.78
DeepLabV3Plus [14]	99.93	81.08	80.58	87.20
UPerNet [15]	99.94	82.85	84.94	89.24
SegFormer [16]	99.95	86.18	86.31	90.81
SegNext [17]	99.96	85.84	85.92	90.57
Ours	**99.97**	**89.93**	**90.09**	**93.33**

Table 1. Segmentation results of different methods on our dataset. The best and second-best results are highlighted in **bold** and underlined.

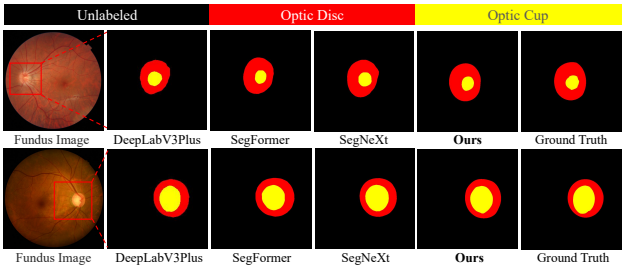

Fig. 2. The segmentation results are shown for each scene in the rows. To better visualize the results, the important regions of each image have been enlarged.

Through this iterative process, the model progressively refines the segmentation mask using fused features, rather than relying on a static fundus image. This approach improves robustness to uncertainty, noise, and modality discrepancies, yielding more accurate and precise boundaries.

3. EXPERIMENTS

3.1. Experiment setup

Dataset: Experiments utilize the GAMMA dataset [12], containing 100 fundus/OCT pairs with glaucoma labels and disc/cup masks. An 80/20 random split is used.
Implementation Details: Experiments are conducted in Pytorch 2.2 on an RTX 4090 GPU for 1×10^5 iterations, learning rate of 1×10^{-5}, and batch size 6.

3.2. Comparison and Analysis

We compare IMO against several segmentation and classification baselines. As shown in Table 1, IMO outperforms other methods in segmentation. Specifically, it achieves 89.93% and 90.09% Dice for the optic disc and cup, respectively, notably surpassing second-best results. Overall, our approach reaches the highest mean Dice of 93.33% on the test set. These results indicate that IMO improves both pixel-level overlap and boundary delineation, which are essential for subsequent glaucoma diagnosis.

To better illustrate the comparison, we provide the visual results of our method alongside other approaches, as shown in Fig. 2. For the first image, our segmentation output is noticeably smoother and more precise than that of the alternative method, with the optic cup region showing natural boundaries and consistent structural details. In the second image, our method achieves higher accuracy and finer delineation, capturing subtle features that are missed or incorrectly segmented by other approaches. In general, the visual comparison demonstrates the superior ability of our method to produce clinically accurate smooth segmentation maps.

Table 2 presents a quantitative comparison of glaucoma grading across different methods. Our approach achieves the highest recall of 79.17% and F1-score of 72.03%, demonstrating strong capability to correctly identify positive cases while maintaining balanced performance. Our precision reaches 81.48%, slightly below ResNet-50 by 0.74%, while our method ties for the highest accuracy of 75.00%. These results indicate that our approach not only performs well in segmentation, but also excels in grading tasks.

3.3. Ablation Study

Framework Ablation: To evaluate the effectiveness of our joint optimization framework, we reconstructed several variant configurations, including single-modality input segmentation-only, and grading-only. As shown in the first three rows of Table 3, using only fundus images (w/o OCT) leads to a decrease in all metrics, with grading Precision decreasing by 11.36%. Optimizing a single task does not outperform joint optimization, demonstrating that our framework effectively leverages the complementary information between segmentation and grading.
CMFA Ablation: To assess the contribution of the CMFA, we remove CMFA and instead directly concatenate features from both modalities. As shown in the fourth row of Table 3, this degradation leads to unstable feature interaction and weakens the modality complementarity. Quantitatively, mDice from 93.33% to 92.72%, and grading Precision drops by 7.24%, indicating that CMFA effectively aligns the features from different modalities and makes it suitable for downstream tasks.
Iterative Refinement Decoder Ablation: As referred to in the fifth row of Table 3, replacing the iterative refinement decoder (IRD) with a single-pass decoder leads to a decrease in mDice of about 0.47%, the impact is relatively minor, with Precision decreasing by only 1.09% relative to the other ablation configurations. This confirms that IRD facilitates progressive feature refinement, leading to more stable segmentation boundaries and improved grade reliability.

4. CONCLUSION

In this paper, we propose the IMO, which leverages complementary information from fundus images and OCT volumes for joint segmentation and grading. The iterative multimodal framework enables mutual supervision between the two tasks, enhancing feature integration and overall performance. Ex-

Methods	Precision↑	Accuracy↑	Recall↑	F1-score↑
ResNet-50 [18]	**82.22**	**75.00**	66.67	66.49
ConvNeXt [19]	79.05	65.00	58.33	58.09
ViT [20]	77.78	70.00	66.67	67.94
BEiTv2 [21]	72.22	65.00	66.67	60.00
EfficientViT [22]	74.24	70.00	70.83	66.25
Ours	81.48	**75.00**	**79.17**	**72.03**

Table 2. Comparison of glaucoma grading performance across different methods.

Method	Disc (Dice↑)	Cup (Dice↑)	mDice↑	Precision↑
w/o OCT	87.38	88.02	91.79	70.12
w/o Grd.	88.93	89.51	92.80	-
w/o Seg.	-	-	-	78.41
w/o CMFA	89.22	88.98	92.72	74.24
w/o IRD	89.03	89.59	92.86	80.39
Ours	**89.93**	**90.09**	**93.33**	**81.48**

Table 3. Ablation studies of IMO. "w/o" denotes without, while "Grd." and "Seg." refer to the grading and segmentation tasks, respectively.

periments on the GAMMA dataset demonstrate strong results in both segmentation and grading. The potential of the framework can be further unlocked with more diverse and well-annotated multimodal data.

5. COMPLIANCE WITH ETHICAL STANDARDS

The use of open-access datasets complied with their license terms and did not require separate ethical approval.

6. REFERENCES

[1] Y. Li *et al.*, "An intelligent and handheld device for early identification of meibomian gland irregularities," in *Ophthalmic Technologies XXXIV*, vol. 12824. SPIE, 2024, pp. 77–81.

[2] Y.-C. Tham *et al.*, "Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis." *Ophthalmology*, vol. 121 11, pp. 2081–90, 2014.

[3] Y. Li *et al.*, "A deep-learning-enabled monitoring system for ocular redness assessment," in *Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, 2023, pp. 1–5.

[4] M. Zhu *et al.*, "Feddm: Federated weakly supervised segmentation via annotation calibration and gradient deconflicting," *IEEE Transactions on Medical Imaging*, vol. 42, pp. 1632–1643, 2023.

[5] S. Cho *et al.*, "Enhanced structure preservation and multi-view approach in unsupervised domain adaptation for optic disc and cup segmentation," *Proc. IEEE Int. Symp. Biomed. Imaging (ISBI)*, pp. 1–5, 2024.

[6] H. lian Liang *et al.*, "Lagnet: Label attention graph networks for ocular disease classification using fundus images," *Proc. IEEE Int. Symp. Biomed. Imaging (ISBI)*, pp. 1–5, 2024.

[7] Z. Cai *et al.*, "Corolla: An efficient multi-modality fusion framework with supervised contrastive learning for glaucoma grading," *Proc. IEEE Int. Symp. Biomed. Imaging (ISBI)*, pp. 1–4, 2022.

[8] Y. Li *et al.*, "Dual-mode imaging system for early detection and monitoring of ocular surface diseases," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 18, pp. 783–798, 2024.

[9] M. Zhu *et al.*, "Dsi-net: Deep synergistic interaction network for joint classification and segmentation with endoscope images," *IEEE Transactions on Medical Imaging*, vol. 40, pp. 3315–3325, 2021.

[10] Z. Wang *et al.*, "Difusionseg: Diffusion-driven semantic segmentation with multi-modal image fusion for enhanced perception," *Knowl. Based Syst.*, vol. 330, p. 114481, 2025.

[11] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *ArXiv*, vol. abs/2006.11239, 2020.

[12] J. Wu, H. Fang, F. Li *et al.*, "Gamma challenge: Glaucoma grading from multi-modality images," *Medical Image Analysis*, vol. 90, p. 102938, 2022.

[13] Z. Zhou *et al.*, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, pp. 1856–1867, 2019.

[14] L.-C. Chen *et al.*, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018.

[15] T. Xiao *et al.*, "Unified perceptual parsing for scene understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018.

[16] E. Xie *et al.*, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Neural Information Processing Systems*, 2021.

[17] M.-H. Guo *et al.*, "Segnext: Rethinking convolutional attention design for semantic segmentation," *ArXiv*, vol. abs/2209.08575, 2022.

[18] K. He *et al.*, "Deep residual learning for image recognition," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 770–778, 2016.

[19] Z. Liu *et al.*, "A convnet for the 2020s," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 11 966–11 976, 2022.

[20] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *ArXiv*, vol. abs/2010.11929, 2020.

[21] Z. Peng *et al.*, "Beit v2: Masked image modeling with vector-quantized visual tokenizers," *ArXiv*, vol. abs/2208.06366, 2022.

[22] X. Liu *et al.*, "Efficientvit: Memory efficient vision transformer with cascaded group attention," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 14 420–14 430, 2023.