RetiDiff: Diffusion-based Synthesis of Retinal OCT Images for Enhanced Segmentation

Sicheng Li[1], Mai Dan[1], Yuhui Chu[1], Jiahui Yu[1], Yunpeng Zhao[1,3], and Pengpeng Zhao[1,2⋆]

[1] Innovation Center for Smart Medical Technologies & Devices, Binjiang Institute of Zhejiang University, Hangzhou, China
[2] College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou, China
[3] Hangzhou Molecular Diagnostics Engineering Research Center, Bioer Technology, Hangzhou, China
zhaopengpeng@zju.edu.cn

Abstract. Optical coherence tomography (OCT) enables detailed visualization and critical segmentation of retinal layers, which is essential for ophthalmological diagnosis. However, the development of automatic segmentation methods has been hindered by limited annotated datasets due to time-consuming manual labeling processes. Therefore, we propose RetiDiff, a three-stage diffusion model-based framework to synthesize realistic annotated OCT retinal images for enhancing segmentation performance. By leveraging the diffusion model, RetiDiff can synthesize diverse and realistic images guided by segmentation masks. To improve synthesis quality and accuracy in pathological regions, we introduce dynamic region masking (DRM), which selectively modifies pathological areas during training. To align the continuous outputs from mask sampling in the diffusion model with discrete segmentation labels, we propose discrete mask clustering (DMC), which converts these outputs into discrete values consistent with the labels. Experimental results show that RetiDiff effectively mitigates data scarcity by synthesizing realistic and diverse annotated OCT retinal images, which substantially enhance retinal layer segmentation performance. Compared to state-of-the-art methods, RetiDiff-synthesized datasets improve the average Dice score by 8.7% across all retinal layers, with a particularly notable increase of up to 53.8% in pathological regions. The code and dataset are publicly available at: https://github.com/MaybeRichard/RetiDiff.

Keywords: Retinal layer segmentation · Data augmentation · Diffusion models · Medical image synthesis

1 Introduction

Optical coherence tomography (OCT) has become a fundamental imaging modality in ophthalmology, offering high-resolution visualization of retinal layers [9].

⋆ Corresponding author

Accurate segmentation of these layers is critical for diagnosing and monitoring ophthalmic diseases such as macular disorders and glaucoma, as it reveals pathological changes through quantitative analysis [3,2]. However, the development of automated segmentation methods is constrained by limited annotated datasets [20,22,10]. This limitation stems from the labor-intensive process of manual annotation, which requires time and expertise from ophthalmologists. Thus, developing cost-effective methods to expand the annotated datasets has become essential for advancing automated retinal analysis.

In recent years, generative models have emerged as a promising solution to address data scarcity in medical imaging [23,15]. While early approaches using generative adversarial networks (GANs) [21,19] showed potential for OCT image synthesis, they suffered from instability and limited diversity. The advent of denoising diffusion probabilistic models (DDPM) [8] addressed many of these issues with more stable training and improved image quality [6,12]. Wu et al. [22] proposed a DDPM-based method that synthesizes retinal images from rough layer sketches and uses knowledge adaptation with pseudo-labels to align synthetic images with their labels. Similarly, Huang et al. [11] developed a transformer-based DDPM for structural label generation paired with a mix-conditional latent diffusion model. Despite these advances, existing DDPM-based methods face two critical limitations: First, they lack fine-grained control over anatomical structures and pathological features during the generation process, leading to limited accuracy in synthetic images. Second, limited annotated datasets restrict the generative models' representation capability, thereby affecting the quality of synthesized images.

Therefore, we propose RetiDiff, a three-stage DDPM-based OCT retinal image synthesis framework. We first pretrain a DDPM using a large amount of unannotated datasets to learn the fundamental representation of retinal images. Second, we train a separate DDPM using segmentation masks from annotated datasets to synthesize diverse masks. Third, we fine-tune the pretrained model with annotated datasets to synthesize high-quality OCT images with segmentation masks as guidance. To address the conflict between generating diverse pathological features and maintaining anatomical consistency in mask-guided models, we propose dynamic region masking (DRM), which selectively masks pathological regions during training to enhance generation of diverse pathological features. To address the misalignment between continuous outputs during the mask sampling process and discrete segmentation labels, we present discrete mask clustering (DMC), which converts these continuous outputs into discrete labels via clustering. Through this framework, our method synthesizes diverse and realistic annotated OCT retinal images, which enhances the performance of automated segmentation, thereby improving diagnostic efficiency and accuracy.

2 Methodology

Diffusion Models. The DDPM we employed consists of two processes: the forward process and the reverse process. Specifically, the forward process gradually

adds Gaussian noise to the input image x_0 over time step t_n, creating a noisy sequence $x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon_t$, where $\epsilon_t \in \mathcal{N}(0, I_n)$ represents the noise at time step t, which is sampled from a standard normal distribution with mean 0 and covariance matrix I_n. The parameter α_t controls the amount of noise added at each time step. The reverse process learns to denoise by training a neural network to predict the noise $\epsilon_\theta(x_t, t)$. Specifically, the reverse process is trained by minimizing the loss function $\mathcal{L} = \mathbb{E}_{x_0,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right]$, where $\epsilon_\theta(x_t, t)$ is the predicted noise at time step t. Through this process, the model is able to synthesize high-quality images from random noise through iterative denoising steps.

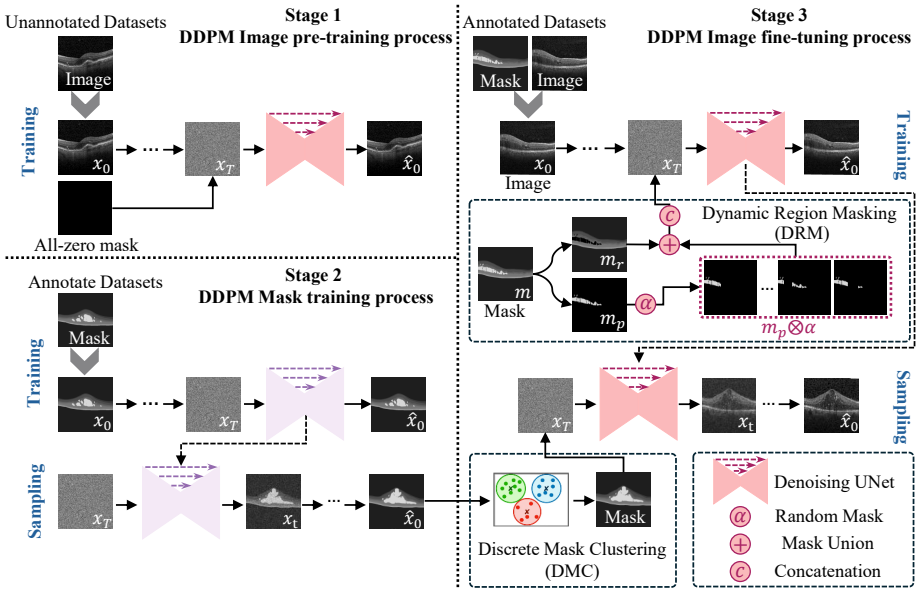

Fig. 1. Overview of our proposed method. The method follows a three-stage approach: (1) Pretraining on unannotated datasets with all-zero mask as condition, (2) Training with annotated mask, and (3) Fine-tuning on the annotated datasets with gound truth mask as condition. Two key methods are introduced: Dynamic Region Masking (DRM) for handling pathological regions, and Discrete Mask Clustering (DMC) to convert continuous mask outputs to discrete labels.

Overall Architecture. Fig.1 shows the overview of RetiDiff, which integrates three stages: two conditional DDPM for image synthesis (stage 1 and stage 3) and an unconditional DDPM for mask generation (stage 2). In stage 1, a large amount of unannotated datasets with all-zero mask as guidance is used to pretrain a DDPM, which improves the models' representation capability. In stage

2, segmentation masks in annotated datasets are used to train a DDPM for mask synthesis. When sampling masks from the trained model in stage 2, we apply DMC to convert continuous diffusion outputs into discrete segmentation labels. In stage 3, we fine-tune the pretrained model from stage 1 with annotated images and masks. During this stage, we employ DRM which separates pathological regions (m_p) from annotated mask (m), applies random masking to pathological region, and recombines them through a mask union operation before using the processed mask as guidance for image synthesis.

2.1 Dynamic Region Masking (DRM)

DRM is a training strategy specifically designed to address challenges in mask-guided diffusion models when dealing with pathological regions. It aims to mitigate the difficulties in synthesizing diverse yet anatomically accurate pathological features, where high variability and complex morphologies often lead to inconsistencies between synthesized images and guidance masks. Given a segmentation mask $m \in \{0, ..., C-1\}^{H \times W}$ with C classes (background, retinal layers, and pathological regions), DRM first identifies pathological regions using intensity thresholds to separate pathological (m_p) and retinal (m_r) masks. We then apply random ablation to m_p via $m_p^{'} = m_p \otimes \alpha$, where $\alpha \in [0,1]^{H \times W}$ is a random mask with elements following uniform(0,1) and probability p of being zero. The final mask $m^{'}$ combines m_r with $m_p^{'}$ through union operation ($m^{'} = m_r \oplus m_p^{'}$), integrating both retinal information and modified pathological regions. This processed mask is concatenated with input x_T during training. Our loss function is defined as:

$$\mathcal{L} = \mathbb{E}_{(x_0, m'), t, \epsilon} \left[\left\| \epsilon - \epsilon_\theta(x_t, t \mid m^{'}) \right\|^2 \right] \tag{1}$$

Through this dynamic masking approach, RetiDiff learns to model the variable characteristics of pathological features like IRF while maintaining their relationship with surrounding tissues, enabling more realistic and diverse OCT image generation.

2.2 Discrete Mask Clustering (DMC)

DMC serves as a post-processing method to reconcile the continuous-valued outputs of diffusion models with the discrete label requirements needed for segmentation tasks. This mismatch often affects boundary definition and downstream model performance. For implementation, we consider a mask $m \in \mathbb{R}^{H \times W}$ with continuous pixel values m_{ij}. We flatten m into a set $\mathcal{X} = \{x_1, ..., x_N\}$ where $N = H \times W$, and apply K-means clustering with $K = C$, where C is the number of segmentation classes. The algorithm minimizes the cost function:

$$J = \sum_{j=1}^{K} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2 \tag{2}$$

where C_j is the set of pixels in cluster j, and μ_j is its centroid. After clustering convergence, we can obtain K cluster centers $\{\mu_1, ..., \mu_K\}$ and map them to discrete label values via $f(\mu_j) = t_j$, where $\{t_1, ..., t_K\}$ correspond to the segmentation classes. The final discrete mask $\widehat{m}$ assigns each pixel the appropriate label: $\widehat{m} = t_{c(i,j)}$, where $c(i, j)$ is the cluster index for pixel (i, j). This process transforms the continuous-valued mask into exactly C discrete classes, ensuring strict alignment with the original segmentation mask space. The resulting discrete masks not only maintain the structural patterns learned by the diffusion model but also guarantee format compatibility with both ground-truth annotations and downstream segmentation algorithms.

3 Experiment

3.1 Dataset and Evaluation Metrics

Datasets. In the first stage, we used the OCT2017 dataset [13], which comprises 84,484 OCT images without segmentation masks for retinal layers. In the second and third stage, we used the training set from the DUKE diabetic macular edema (DME) dataset [5]. This dataset includes 110 OCT B-scan images with corresponding retinal segmentation masks, split into 66 pairs for training, 22 pairs for validation, and 22 pairs for testing. For the downstream segmentation task, we tested segmentation performance using the DUKE DME test set.

Evaluation Metrics. We adopted the fréchet inception distance (FID) and learned perceptual image patch similarity (LPIPS) for generative quality assessment, and Dice score (DSC) and pixel accuracy (PA) for segmentation performance evaluation.

Model Implementation. All experiments were conducted on an NVIDIA RTX 4090 GPU. The model was trained in three stages: pretraining for 10,000 epochs in stage 1, followed by 3,000 epochs for mask synthesis (stage 2) and 3,000 epochs for image fine-tuning (stage 3). We used AdamW optimizer [14] with learning rate 1e-5 and batch size 2. All images were normalized to [-1, 1] range and resized to 480×480 pixels.

3.2 Synthesis Results and Ablation Study

We conducted qualitative and quantitative comparisons between RetiDiff and other generative models, as well as ablation studies on each RetiDiff component. As demonstrated in Fig. 2 (A), RetiDiff achieved better mask-guided image synthesis with higher anatomical consistency, while both LDM [16] (DDPM with autoencoder) and Retree [1] (DDPM with local self-attention mechanism and multi-stage conditional concatenation) exhibited misalignment error in IRF and other retinal regions, failing to maintain accurate correspondence with the guidance masks. Table 2 compares the changes in segmentation performance after

training with different datasets, unlike LDM, which relied on the variational autoencoder to encode data into latent space. Retree used a more traditional DDPM pipeline, achieving better FID and LPIPS scores. This improvement likely stemmed from DDPM avoiding information loss during latent compression and more efficiently integrating conditional mask information directly in pixel space. In ablation study, we evaluated the impact of each proposed component in RetiDiff. With the gradual addition of different components, the the quality of the synthesized images improved significantly. The pretraining process was crucial in helping the model establish a fundamental understanding of OCT image structures, improving FID by 34.59%. With the addition of DRM strategy, the model's ability to characterize IRF regions was enhanced, further improving FID by 21.81%. Overall, RetiDiff outperformed existing methods in both FID and LPIPS metrics, demonstrating the effectiveness and complementarity of our proposed components in generating high-quality, anatomically consistent OCT images.

Table 1. Quantitative comparison of state-of-the-art generative models and ablation stydy of Retidiff. (🡅: Higher is better, 🡇: Lower is better)

Method	Pretrained	DRM	DMC	FID 🡇	LPIPS 🡅
Retree	-/-	-/-	-/-	62.9871	0.4189
LDM	-/-	-/-	-/-	84.6451	0.3225
RetiDiff	✖	✖	✖	68.5206	0.3832
	✔	✖	✖	44.8165	0.4156
	✔	✔	✖	35.0441	0.4331
	✔	✔	✔	**31.7257**	**0.4474**

3.3 Segmentation Results

We further evaluated the synthetic dataset in retinal layer segmentation tasks. In the experiments, we created three training data groups: (1) R/S (66/0), where "R" stands for real and "S" stands for synthetic. This group comprises 66 pairs of images, each consisting of a real annotated image from the DUKE DME training set and its corresponding segmentation mask, with no synthetic images included. (2) R/S (0/66): This group has 66 synthetic annotated images synthesized by RetiDiff using masks from the DUKE DME training set as conditions. This ensures a direct comparison with real annotated images. (3) R/S (0/1000): This group contains 1000 synthetic annotated images synthesized by RetiDiff, where the masks from the stage 2 sampling process are used as conditions. These three groups of training data were used to train four segmentation models including UNet [17], YNet [7], ReLayNet [18], and GDNet [4]. All models were tested using the DUKE DME test set and results are shown in Fig. 2 (B) and Table 2.

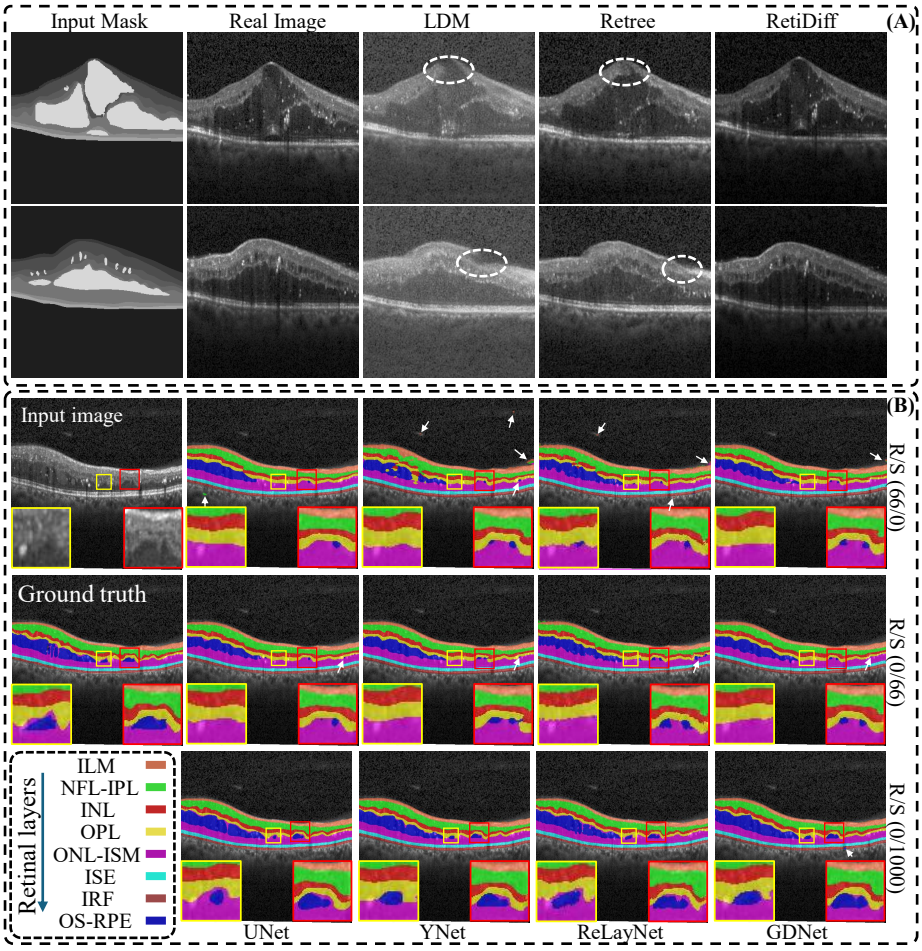

Fig. 2. Comparison of synthesis results from different methods (A), and comparison of OCT retinal image segmentation results across different datasets and models (B). White arrows indicate notable segmentation errors, with rectangular areas enlarged for comparison. White dashed lines mark inaccurate regions in the synthesized images.

As shown in Fig. 2 (B), white arrows highlighted segmentation errors in the first and second rows. The first row, trained on R/S (66/0), had the most errors, primarily in the background region. The second row, trained on R/S (0/66), showed errors mainly in the IRF regions. In the small IRF regions marked in red and yellow, the first two rows significantly differ from the ground-truth. In contrast, the third row trained on R/S (0/1000), not only eliminated the segmentation errors in the background and retinal regions present in the first two rows, but also significantly improved the segmentation accuracy in the IRF region. These observations were supported by quantitative results in Table 1, which

evaluated three key retinal layers: NFL-IPL (nerve fiber layer to inner plexiform layer), ONL-ISM (outer nuclear layer to inner segment myocardium), and IRF, using the DUKE DME test set. The R/S (66/0) and R/S (0/66) datasets gave similar average metrics, with R/S (66/0) slightly better. This indicated that real dataset had a small edge over synthetic ones when using the same masks, showing the synthetic images were close but not fully equal to real dataset in segmentation tasks. However, in the R/S (0/1000) dataset, the significant increase in the quantity and diversity of annotated datasets led to a notable improvement in the performance of multiple segmentation models. Notably, YNet showed the largest gain compared to training on real data alone: the DSC for the IRF region rose by 53.8%, PA by 33.8%, and the average DSC and PA improved by 8.7% and 3.2%, respectively. Qualitative and quantitative experimental results show that our method improves the segmentation accuracy of multiple models by synthesizing a large number of different retinal OCT annotation samples.

Table 2. Quantitative comparison of segmentation methods across different training datasets. (↑: Higher is better)

Real/ Synthesis	Model	NFL-IPL		ONL-ISM		IRF		Mean	
		DSC ↑	PA ↑	DSC↑	PA ↑	DSC ↑	PA ↑	DSC ↑	PA ↑
R/S (66/0)	UNet	0.9048	0.8823	0.9026	0.9015	0.5825	0.5804	0.8221	0.8732
	YNet	0.8769	0.8655	0.8909	0.8352	0.5622	0.5901	0.8038	0.8849
	ReLayNet	0.9085	0.8915	0.9043	0.8985	0.5819	0.5910	0.8271	0.8754
	GDNet	0.9011	0.8764	0.9042	0.8901	0.6410	0.6059	0.8391	0.8875
R/S (0/66)	UNet	0.8964	0.8833	0.9013	0.9004	0.5859	0.5680	0.8145	0.8643
	YNet	0.8976	0.8725	0.9023	0.8767	0.5903	0.6607	0.8118	0.8489
	ReLayNet	0.8788	0.8789	0.8742	0.8731	0.5732	0.6953	0.8005	0.8489
	GDNet	0.9051	0.8786	0.9005	0.8728	0.6238	0.6543	0.8234	0.8883
R/S (0/1000)	UNet	**0.9389**	**0.9262**	**0.9347**	**0.9383**	**0.7511**	**0.8408**	**0.8629**	**0.8974**
	YNet	**0.9411**	**0.9259**	**0.9306**	**0.9130**	**0.8649**	**0.7901**	**0.8738**	**0.9135**
	ReLayNet	**0.9365**	**0.9347**	**0.9343**	**0.9371**	**0.8010**	**0.8339**	**0.8663**	**0.8772**
	GDNet	**0.9393**	**0.9215**	**0.9268**	**0.8911**	**0.7947**	**0.9215**	**0.8579**	**0.8991**

4 Conclusions

This paper presents RetiDiff, a three-stage DDPM for synthesizing high-quality, anatomically consistent annotated OCT retinal images. To address the challenge of generating accurate annotated images with limited labeled data, we pretrain the model on a large amount of unannotated datasets and incorporate the DRM strategy along with the DMC post-processing method, which significantly improving both image quality and downstream segmentation performance. Experimental results demonstrate that our method outperforms existing approaches in multiple metrics, with a remarkable 53.8% increase in Dice score for lesion regions. Future work will extend this method to other pathological imaging modal-

ities, providing cost-effective annotated datasets to support various diagnostic applications.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (62305289), Hangzhou Science and Technology Bureau (TD2023018) and Zhejiang Province Postdoctoral Research Selection Funding Project (ZJ2022008).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alimanov, A., Islam, M.B.: Denoising diffusion probabilistic model for retinal image generation and segmentation. In: IEEE International Conference on Computational Photography (ICCP). pp. 1–12. IEEE (2023)
2. Borrelli, E., Sarraf, D., Freund, K.B., Sadda, S.R.: Oct angiography and evaluation of the choroid and choroidal vascular disorders. Progress in Retinal and Eye Research **67**, 30–55 (2018)
3. Bussel, I.I., Wollstein, G., Schuman, J.S.: Oct for glaucoma diagnosis, screening and detection of glaucoma progression. British Journal of Ophthalmology **98**(Suppl 2), ii15–ii19 (2014)
4. Cao, G., Zhou, Z., Wu, Y., Peng, Z., Yan, R., Zhang, Y., Jiang, B.: Gcn-enhanced spatial-spectral dual-encoder network for simultaneous segmentation of retinal layers and fluid in oct images. Biomedical Signal Processing and Control **98**, 106702 (2024)
5. Chiu, S.J., Allingham, M.J., Mettu, P.S., Cousins, S.W., Izatt, J.A., Farsiu, S.: Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema. Biomedical Optics Express **6**(4), 1172–1194 (2015)
6. Du, Y., Jiang, Y., Tan, S., Wu, X., Dou, Q., Li, Z., Li, G., Wan, X.: Arsdm: colonoscopy images synthesis with adaptive refinement semantic diffusion models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 339–349. Springer (2023)
7. Farshad, A., Yeganeh, Y., Gehlbach, P., Navab, N.: Y-net: A spatiospectral dual-encoder network for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 582–592. Springer (2022)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
9. Huang, D., Swanson, E.A., Lin, C.P., Schuman, J.S., Stinson, W.G., Chang, W., Hee, M.R., Flotte, T., Gregory, K., Puliafito, C.A., et al.: Optical coherence tomography. Science **254**(5035), 1178–1181 (1991)
10. Huang, K., Ma, X., Zhang, Y., Su, N., Yuan, S., Liu, Y., Chen, Q., Fu, H.: Memory-efficient high-resolution oct volume synthesis with cascaded amortized latent diffusion models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 478–487. Springer (2024)
11. Huang, K., Ma, X., Zhang, Z., Zhang, Y., Yuan, S., Fu, H., Chen, Q.: Diverse data generation for retinal layer segmentation with potential structure modelling. IEEE Transactions on Medical Imaging (2024)

12. Jiang, L., Mao, Y., Wang, X., Chen, X., Li, C.: Cola-diff: Conditional latent diffusion model for multi-modal mri synthesis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 398–408. Springer (2023)
13. Kermany, D.: Labeled optical coherence tomography (oct) and chest x-ray images for classification. Mendeley Data (2018)
14. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
15. Oh, H.J., Jeong, W.K.: Diffmix: Diffusion model-based data synthesis for nuclei segmentation and classification in imbalanced pathology image datasets. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 337–345. Springer (2023)
16. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (2022)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 234–241. Springer (2015)
18. Roy, A.G., Conjeti, S., Karri, S.P.K., Sheet, D., Katouzian, A., Wachinger, C., Navab, N.: Relaynet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks. Biomedical Optics Express **8**(8), 3627–3642 (2017)
19. Tajmirriahi, M., Kafieh, R., Amini, Z., Lakshminarayanan, V.: A dual-discriminator fourier acquisitive gan for generating retinal optical coherence tomography images. IEEE Transactions on Instrumentation and Measurement **71**, 1–8 (2022)
20. Upadhyay, A.K., Bhandari, A.K.: Advances in deep learning models for resolving medical image segmentation data scarcity problem: A topical review. Archives of Computational Methods in Engineering **31**(3), 1701–1719 (2024)
21. Vidal, P.L., de Moura, J., Novo, J., Penedo, M.G., Ortega, M.: Image-to-image translation with generative adversarial networks via retinal masks for realistic optical coherence tomography imaging of diabetic macular edema disorders. Biomedical Signal Processing and Control **79**, 104098 (2023)
22. Wu, Y., He, W., Eschweiler, D., Dou, N., Fan, Z., Mi, S., Walter, P., Stegmaier, J.: Retinal oct synthesis with denoising diffusion probabilistic models for layer segmentation. In: IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2024)
23. Ye, J., Ni, H., Jin, P., Huang, S.X., Xue, Y.: Synthetic augmentation with large-scale unconditional pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 754–764. Springer (2023)