Propagating Structural Guidance: Synthesizing Fluorescein Angiography from Fundus Images and Sparse OCT Scans

Tengfei Ma[1,2], Ruiqi Wu[1,2], Chenran Zhang[1,2], Ye Geng[3], Na Su[5], Xiangyuan Duanmu[3], Tao Zhou[4], Yi Zhou[1,2(✉)], Wen Fan[5(✉)]

School of Computer Science and Engineering, Southeast University, China[1]
Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications, Ministry of Education, Nanjing, China[2]
Tianyuan Honors School, Nanjing Medical University, China[3]
Nanjing University of Science and Technology, Nanjing, China[4]
Department of Ophthalmology, The First Affiliated Hospital of Nanjing Medical University, Nanjing, China[5]
{matengfei_1013@163.com, yizhou.szcn@gmail.com}

Abstract. Fundus fluorescein angiography (FFA) is critical for assessing retinal vascular abnormalities, but its acquisition is invasive and not always feasible. In contrast, color fundus photography (CFP) is non-invasive and widely accessible, which has motivated studies on CFP-to-FFA synthesis. However, prior works rely solely on CFP surface texture, fundamentally limiting the ability to reconstruct functional vascular information and subtle pathological changes. To address this, we propose a novel framework that synthesizes FFA from CFP with structural guidance provided by optical coherence tomography (OCT). We construct a multi-modal retinal imaging dataset with paired CFP, FFA, and OCT from 3,676 patient eyes—the first tri-modally aligned dataset in retinal imaging. To bridge the spatial gap between OCT and fundus modalities, we propose a Spatially Aligned Cross-Modal Fusion (**SACMF**) module that projects depth-resolved OCT features onto the fundus plane and injects them into the CFP encoder via adaptive layer normalization. Beyond feature fusion, we further introduce Token-wise Cross-Modality Alignment (**TCMA**), a token-level contrastive learning strategy that explicitly aligns CFP and FFA representations at corresponding spatial positions. Our method achieves superior synthesis performance compared to state-of-the-art methods. Moreover, extensive experiments demonstrate that the FFA images synthesized by our approach bring greater improvements in downstream disease diagnosis performance than existing methods, highlighting the clinical potential of our approach as a non-invasive decision-support tool in routine workflows. The code is available at https://github.com/while-plus/OCT-guide-FFA-Syn.

Keywords: Multimodal-driven FFA Synthesis · OCT Structural Guidance · Token-wise Contrastive Learning

1 Introduction

Fundus fluorescein angiography (FFA) plays a critical role in retinal disease assessment by visualizing vascular perfusion, leakage, and pathological microcirculation changes, essential for diagnosing diabetic retinopathy and retinal vein occlusion [14]. However, FFA is invasive, requiring intravenous dye injection, which limits routine and repeated use. In contrast, color fundus photography (CFP) is non-invasive and easily acquired [26], suitable for large-scale screening and longitudinal follow-up. Nevertheless, CFP primarily captures retinal surface appearance, lacking explicit information about vascular permeability and dynamic blood flow [14]. This limitation restricts its ability to reflect functional vascular abnormalities central to many retinal pathologies. This gap motivates computational approaches that infer vascular-related information from CFP, aiming to approximate FFA without direct angiographic acquisition.

Despite recent advances in generative modeling for FFA synthesis, most existing works [24,8,12,11] rely solely on CFP input. Consequently, these approaches are constrained by CFP's limited information, as it does not encode vascular leakage or depth-resolved tissue properties. This fundamentally restricts the upper bound of CFP-to-FFA synthesis, often producing images lacking fine-grained vascular details. More critically, networks may generate fluorescence patterns based on superficial texture correlations rather than true vascular structures. Such artifacts underscore the intrinsic limitations of CFP-only approaches and severely constrain the clinical interpretability of synthesized FFA.

To address these limitations, our study is the first to incorporate optical coherence tomography (OCT) as complementary structural guidance for FFA generation, bridging the modality gap between fundus photography and angiography. Similar to CFP, OCT is non-invasive and safely acquired when FFA is contraindicated, making it a practical auxiliary modality. In intraoperative scenarios, only CFP-like views and OCT are typically available, as FFA is invasive and procedurally impractical. Therefore, synthesizing FFA from CFP and OCT is valuable for surgical decision-making requiring real-time vascular assessment [3]. Importantly, OCT provides high-resolution, layer-resolved structural information that CFP cannot capture. Structural abnormalities such as retinal thickening or abnormal reflectivity patterns in OCT are often associated with fluorescence leakage or accumulation in FFA [18]. Incorporating OCT introduces anatomically grounded cues, reducing the gap between surface appearance and underlying vascular pathology, enabling more reliable fluorescence pattern modeling. Considering clinical workflows, only limited representative two-dimensional B-scan slices are typically retained instead of full volumetric scans. Accordingly, we focus on sparse B-scans as input, ensuring practical feasibility while reducing data acquisition burden and storage requirements.

Our major contributions are highlighted as: 1) We are the first to investigate CFP-to-FFA synthesis in a multi-modal pipeline by incorporating OCT as auxiliary structural guidance, constructing the first paired tri-modal retinal dataset with 3,676 aligned CFP, FFA, and OCT cases. **2)** We propose an OCT-guided feature modulation module that projects depth-aware OCT representa-

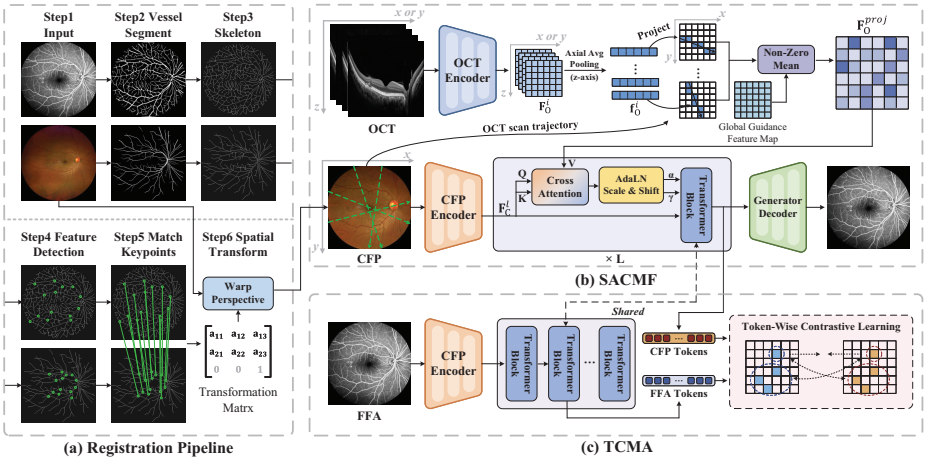

Fig. 1. Registration Pipeline and Overall Architecture. (a) Registration pipeline. Establishes spatial correspondence between CFP and FFA. (b) SACMF. Bridges the viewpoint gap between OCT and CFP and propagates structural guidance across the fundus plane. Green dashed lines indicate OCT scan trajectories. (c) TCMA. Enforces fine-grained semantic consistency between CFP and FFA.

tions onto the fundus plane and injects them into the CFP encoder via adaptive normalization. **3)** We introduce a token-wise cross-modality contrastive learning strategy that enforces CFP-FFA alignment at matched spatial positions. **4)** Extensive experiments demonstrate that our approach generates FFA images with higher visual fidelity, achieving superior downstream diagnostic performance, and exhibits greater clinical utility compared to baselines.

2 Methods

2.1 Overview

Data Curation. Each case was collected from real-world clinical practice at a hospital and includes one CFP, one FFA, and multiple clinician-selected OCT slices considered diagnostically informative. As illustrated in Fig. 1(a), CFP and FFA are first spatially registered to enable cross-modal correspondence. Vessel segmentation is performed for both modalities. CFP vessels are segmented using Automorph [29], while FFA vessels are segmented using a model trained with a style loss [27] on the CHASE [20], DRIVE [22], HRF [1] and STARE [7] datasets. From the segmented vessels, skeletons [23] are extracted to identify endpoints and bifurcation points. ORB descriptors [21] are computed at these keypoints to characterize local vessel structures. Feature correspondences are then matched to estimate a homography matrix. The estimated transformation is applied via perspective warping to align both the vessel masks and the fundus images. Cases with a Dice coefficient [19] below 0.4 between the aligned vessel

masks are excluded. Thus, our final curated dataset contains 3,676 cases, with an average of 2.88 OCT slices per case.

Overall Architecture. Although diffusion models [6] have recently achieved remarkable performance in natural image generation, our task involves paired cross-modal translation with limited medical data and strict structural consistency constraints. In this scenario, conditional GANs show promise due to their strong pixel-level supervision and training efficiency. Therefore, we adopt a Generative Adversarial Network (GAN) [13] as the backbone for CFP-to-FFA synthesis. The framework consists of three main components: (1) a generator with dual encoders for CFP and OCT feature extraction, (2) the proposed Spatial Aligned Cross-Modal Fusion (SACMF, Fig. 1(b)) module for structurally consistent multi-modal fusion, (3) a Token-wise Cross-Modality Alignment (TCMA, Fig. 1(c)) module that enforces fine-grained semantic consistency during training. A discriminator is further introduced to encourage the realism of the synthesized FFA images under adversarial supervision.

2.2 Spatial Aligned Cross-Modal Fusion (SACMF)

The proposed SACMF module consists of two sequential steps: (1) trajectory-aware en-face projection and (2) CFP-guided structural propagation. The first step aims to resolve the viewpoint discrepancy between cross-sectional OCT B-scans and en-face fundus images by constructing a spatially aligned structural representation on the 2D fundus plane. The second step propagates sparse OCT-derived structural cues to CFP regions without direct OCT coverage by leveraging appearance similarity among CFP feature tokens, thereby enabling global structural guidance for FFA synthesis.

Trajectory-Aware En-face Projection. Due to the viewpoint discrepancy between OCT B-scans, which capture retinal anatomy along the x-z or y-z plane [16], and CFP/FFA, which represent en-face appearance in the x-y plane, directly fusing the two modalities introduces spatial inconsistency.

Given N OCT B-scan slices $\mathbf{X}_{\mathrm{O}}^{i}$ $(i = 1 \cdots N)$ co-registered with the fundus space, each slice is processed by a frozen pretrained OCT encoder $\mathbf{E}_{\mathrm{O}}$, producing feature maps $\mathbf{F}_{\mathrm{O}}^{i} \in \mathbb{R}^{C\times H\times W}$. Average pooling is applied along the axial (z) direction to obtain strip-like representations $\mathbf{f}_{\mathrm{O}}^{i} \in \mathbb{R}^{C\times 1\times W}$. According to the known scanning trajectories, each $\mathbf{f}_{\mathrm{O}}^{i}$ is embedded into a zero-initialized 2D feature map $\mathbf{M}_i \in \mathbb{R}^{C\times H\times W}$ at its corresponding location. All embedded maps $\{\mathbf{M}_i\}_{i=1}^{N}$ are aggregated with a learnable global guidance map $\mathbf{G} \in \mathbb{R}^{C\times H\times W}$ via non-zero-aware averaging:

$$\mathbf{F}_{\mathrm{O}}^{proj} = \frac{\sum_{i=1}^{N} \mathbf{M}_i + \mathbf{G}}{\sum_{i=1}^{N} \mathbb{I}[\mathbf{M}_i \neq \mathbf{0}] + \mathbb{I}[\mathbf{G} \neq \mathbf{0}]}, \tag{1}$$

where $\mathbb{I}[\cdot]$ denotes an element-wise indicator function.

CFP-Guided Structural Propagation. Although $\mathbf{F}_{\mathrm{O}}^{proj}$ provides spatially aligned structural anchors, large regions in CFP remain without direct OCT

measurements. To propagate structural cues beyond sparse trajectories, we exploit long-range appearance similarity within CFP features. Given the CFP input $\mathbf{X}_{\mathrm{C}}$, the l-th encoder layer produces $\mathbf{F}_{\mathrm{C}}^{l}$. The query and key are derived from CFP features to measure appearance compatibility, while the projected OCT feature serves as the value to inject structural information. Accordingly, structural guidance is propagated as follows:

$$\mathbf{F}_{\mathrm{prop}}^{l} = \mathrm{softmax}\left(\frac{Q(\mathbf{F}_{\mathrm{C}}^{l})K(\mathbf{F}_{\mathrm{C}}^{l})^{\top}}{\sqrt{d}}\right)V(\mathbf{F}_{\mathrm{O}}^{proj}). \tag{2}$$

This allows OCT-uncovered pixels to retrieve structural cues via CFP-based similarity. The propagated feature $\mathbf{F}_{\mathrm{prop}}^{l}$ is injected into $\mathbf{F}_{\mathrm{C}}^{l}$ via adaptive layer normalization for structure-aware FFA synthesis.

2.3 Token-wise Cross-Modality Alignment (TCMA)

For each input CFP–FFA pair, both images are first processed by a shared CFP encoder, followed by a stack of L transformer blocks. The l-th block outputs CFP and FFA token representations $\mathbf{T}_{\mathrm{C}}, \mathbf{T}_{\mathrm{F}} \in \mathbb{R}^{K\times d}$, where K and d denote the number of spatial tokens and feature dimension. Notably, the CFP tokens $\mathbf{T}_{\mathrm{C}}$ are adaptively modulated using OCT-derived information via AdaLN, while the FFA tokens $\mathbf{T}_{\mathrm{F}}$ remain unmodulated. This asymmetric design allows CFP tokens, conditioned on OCT guidance, to consistently reflect the underlying FFA signals, while keeping FFA tokens unmodulated to prevent redundant or biased information. To further promote cross-modality consistency, we adopt a symmetric InfoNCE objective to enforce bidirectional alignment between CFP and FFA tokens. Tokens at the same spatial location are treated as positive pairs, and all other tokens as negatives. Specifically, the InfoNCE loss is defined as:

$$\mathcal{L}_{\mathrm{InfoNCE}}(\mathbf{q}, \mathbf{k}^{+}, \{\mathbf{k}^{j}\}_{j=1}^{K}) = -\log\frac{\exp(\mathrm{sim}(\mathbf{q}, \mathbf{k}^{+})/\tau)}{\sum_{j=1}^{K}\exp(\mathrm{sim}(\mathbf{q}, \mathbf{k}^{j})/\tau)}, \tag{3}$$

where $\mathbf{q}$ is a query token, $\mathbf{k}^{+}$ is its positive key, and $\{\mathbf{k}^{j}\}_{j=1}^{K}$ denotes the set of keys for contrastive comparison. $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is a temperature hyperparameter. Accordingly, the token-wise contrastive loss is formulated as:

$$\begin{aligned}\mathcal{L}_{\mathrm{tok\_cont}}(T_C, T_F) = \frac{1}{K}\sum_{i=1}^{K}\Big(&\mathcal{L}_{\mathrm{InfoNCE}}(\mathbf{T}_{\mathrm{C}}^{i}, \mathbf{T}_{\mathrm{F}}^{i}, \{\mathbf{T}_{\mathrm{F}}^{j}\}) \\ &+ \mathcal{L}_{\mathrm{InfoNCE}}(\mathbf{T}_{\mathrm{F}}^{i}, \mathbf{T}_{\mathrm{C}}^{i}, \{\mathbf{T}_{\mathrm{C}}^{j}\})\Big).\end{aligned} \tag{4}$$

Applying this supervision across multiple layers aligns structural and functional representations while preserving local details in the fundus image. Notably, TCMA is activated only during training with real FFA tokens for contrastive supervision; during inference, the model relies solely on CFP inputs.

Overall Training Objective. In addition to the proposed token-wise contrastive loss, we follow the baseline setting in [13], incorporating registration loss $\mathcal{L}_{\text{Reg}}$ and adversarial loss $\mathcal{L}_{\text{GAN}}$ to ensure stable training and realistic synthesis. Consequently, the final objective is defined as:

$$L = \mathcal{L}_{\text{GAN}} + \mathcal{L}_{\text{Reg}} + \beta \mathcal{L}_{\text{tok\_cont}}, \tag{5}$$

where β balances the contribution of the contrastive supervision.

3 Experiments

Implementation Details. All images were resized to 512×512. The synthesis model trained for 160 epochs using Adam with a learning rate of 1e-5 and a batch size of 4. The dataset split was 60% training and 40% testing. The weighting coefficient β in Eq. 5 was set to 0.1. All experiments were conducted on four NVIDIA GeForce RTX 4090 GPUs.

To evaluate diagnostic utility, we performed a downstream classification task on three prevalent diseases: diabetic retinopathy, retinitis, and choroidal disorders. Using 75% of these samples for training and the rest for testing, we employed a ResNet18 backbone optimized by Adam at a learning rate of 5e-4 for 50 epochs. This setup assesses whether synthesized images enhance disease diagnosis performance beyond standard image quality metrics.

Evaluation Criteria. For image quality evaluation, we report PSNR [9], SSIM [25], LPIPS [28], and FID [5] to assess pixel fidelity, structural consistency, perceptual similarity, and distribution alignment. Classification performance is measured by Precision (Pre), Recall (Rec), and F1-score (F1). The lower bound takes OCT and CFP as input, while the upper bound additionally incorporates real FFA. All other methods follow the upper-bound setting but replace real FFA with synthesized FFA for classification. Semantic consistency is evaluated via cosine similarity (Cos Sim) between feature embeddings of synthesized and real images, extracted by an encoder trained on FFA-IR [17] [4].

Quantitative Comparisons. Table 1 reports quantitative comparisons with existing fundus translation methods [12,11] and representative natural and medical image generation models [15,10,2,13]. **Image Quality.** Our method achieves the best performance across PSNR, SSIM, FID, and LPIPS, demonstrating superior structural fidelity and perceptual realism. The consistent gains over both fundus-specific and general generation models validate the benefit of incorporating OCT as structural guidance for angiographic synthesis. **Diagnosis & Semantic Consistency.** It also shows that using synthesized FFA improves the lower bound of multi-disease classification. Our approach attains the highest Precision, Recall, and F1-score, narrowing the gap toward the upper bound with real FFA. Moreover, it achieves the highest cosine similarity in semantic embedding space, indicating stronger alignment with real angiographic representations.

Table 1. Comparison of different methods. The best results are highlighted in **bold**.

Method	Image Quality				Diagnosis(%)			Semantic(%)
	PSNR↑	SSIM↑	FID↓	LPIPS↓	Pre↑	Rec↑	F1↑	Cos Sim↑
Lower bound	-	-	-	-	45.59	42.66	43.16	-
Upper bound	-	-	-	-	54.18	53.66	53.80	-
AAGAN [12]	15.48	0.4370	11.23	0.2224	47.40	45.61	46.23	76.59
BBDM [15]	13.77	0.3167	16.23	0.2816	45.28	44.32	44.57	59.01
FAGAN [11]	15.54	0.4276	10.22	0.2146	48.11	47.68	47.76	77.30
P2P [10]	15.22	0.4316	13.92	0.2294	48.89	46.31	47.16	77.84
ResVit [2]	15.59	0.4325	12.89	0.2172	48.96	47.30	47.97	77.28
RegGAN [13]	15.36	0.4223	12.74	0.2135	46.58	45.44	45.84	76.58
Ours	**15.87**	**0.4502**	**10.05**	**0.2009**	**52.83**	**50.61**	**51.49**	**79.13**

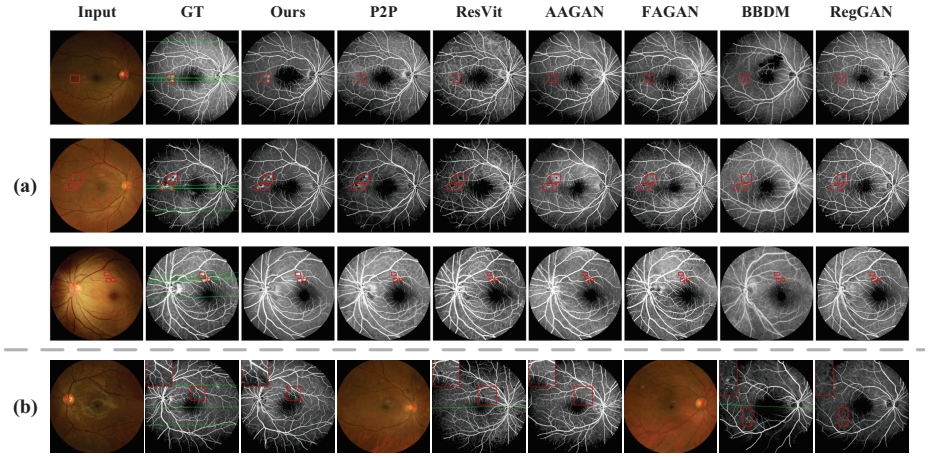

Fig. 2. (a) Qualitative comparison of FFA synthesis results across different methods. (b) Failure cases and limitations of the proposed method. Areas in red boxes denote subtle leakage. Green lines indicate OCT scan trajectories.

Qualitative Comparison. As shown in Fig. 2, CFP-only methods often miss subtle leakage regions (red boxes), particularly small hyperfluorescent spots. These approaches tend to produce over-smoothed responses, resulting in attenuated or absent leakage signals. In contrast, our OCT-guided model better preserves fine-grained leakage patterns, indicating improved pathology-aware synthesis. However, as shown in Fig. 2(b), we also observe certain limitations. Vessel boundaries may appear blurred, and the model struggles to recover very small or widely distributed leakage regions, resulting in incomplete synthesis.

Table 2. Ablation study on our proposed SACMF and TCMA modules. The best results are highlighted in **bold**.

SACMF	TCMA	Image Quality				Diagnosis (%)			Semantic(%)
		PSNR↑	SSIM↑	FID↓	LPIPS↓	Pre↑	Rec↑	F1↑	Cos Sim↑
✗	✗	15.36	0.4223	12.74	0.2135	46.58	45.44	45.84	76.58
✓	✗	15.51	0.4349	10.42	0.2069	49.10	47.33	47.93	78.53
✓	✓	**15.87**	**0.4502**	**10.05**	**0.2009**	**52.83**	**50.61**	**51.49**	**79.13**

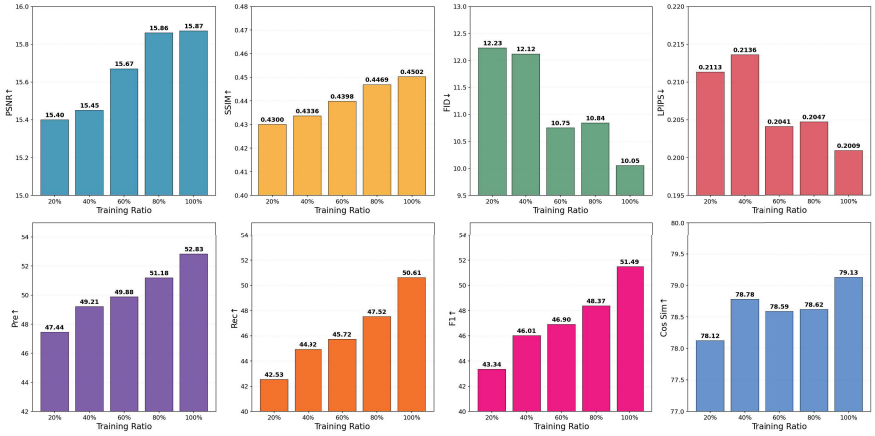

Fig. 3. Impact of different OCT training ratios (20%–100%).

Ablation Studies. We conduct ablation studies to assess the contributions of SACMF and TCMA (Table 2). Adding SACMF improves perceptual quality (FID 12.74→10.42, LPIPS 0.2135→0.2069) and increases F1-score from 45.84% to 47.93%, demonstrating effective incorporation of OCT structural cues. Further introducing TCMA yields consistent gains across all metrics, with PSNR/SSIM reaching 15.87/0.4502 and FID decreasing to 10.05. Notably, F1 rises to 51.49% and cosine similarity to 79.13%, indicating enhanced diagnostic performance and semantic alignment. These results suggest that TCMA enhances semantic-level cross-modal alignment beyond conventional pixel-wise reconstruction objectives.

Impact of OCT Training Ratio. We evaluate the impact of varying OCT training ratios (20%–100%) while keeping CFP unchanged. As shown in Fig. 3, as the OCT proportion increases, all metrics exhibit consistent improvement trends. These metric trajectories collectively confirm that denser OCT supervision strengthens structural guidance propagation and leads to more reliable cross-modal generation.

4 Conclusion

In this work, we present the first multi-modal CFP-to-FFA synthesis framework incorporating OCT as structural guidance. A tri-modal dataset with 3,676 paired CFP, FFA, and sparse OCT scans enables anatomically grounded learning. The proposed SACMF module projects depth-resolved OCT features onto the fundus plane for structural modulation, while TCMA enforces token-level CFP-FFA alignment. Experiments demonstrate superior image quality and discriminative representations for disease diagnosis. Integrating OCT alleviates the information gap in CFP-only translation, offering a practical direction for non-invasive angiography approximation in clinical settings.

Acknowledgement. This work was partially supported by the National Natural Science Foundation of China (Grant No 62476054, Grant No 82401284, and Grant No 62576153).

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

References

1. Budai, A., Bock, R., Maier, A., Hornegger, J., Michelson, G.: Robust vessel segmentation in fundus images. International journal of biomedical imaging **2013**(1), 154860 (2013)
2. Dalmaz, O., Yurt, M., Çukur, T.: Resvit: Residual vision transformers for multimodal medical image synthesis. IEEE Transactions on Medical Imaging **41**(10), 2598–2614 (2022)
3. Ehlers, J.P., Tao, Y.K., Srivastava, S.K.: The value of intraoperative optical coherence tomography imaging in vitreoretinal surgery. Current opinion in ophthalmology **25**(3), 221–227 (2014)
4. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 7514–7528 (2021)
5. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
6. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
7. Hoover, A., Kouznetsova, V., Goldbaum, M.: Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Transactions on Medical Imaging **19**(3), 203–210 (2000). `https://doi.org/10.1109/42.845178`
8. Huang, K., Li, M., Yu, J., Miao, J., Hu, Z., Yuan, S., Chen, Q.: Lesion-aware generative adversarial networks for color fundus image to fundus fluorescein angiography translation. Computer Methods and Programs in Biomedicine **229**, 107306 (2023)
9. Huynh-Thu, Q., Ghanbari, M.: Scope of validity of psnr in image/video quality assessment. Electronics letters **44**(13), 800–801 (2008)

10. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on (2017)
11. Kamran, S.A., Fariha Hossain, K., Tavakkoli, A., Zuckerbrod, S., Baker, S.A., Sanders, K.M.: Fundus2angio: a conditional gan architecture for generating fluorescein angiography images from retinal fundus photography. In: International Symposium on Visual Computing. pp. 125–138. Springer (2020)
12. Kamran, S.A., Hossain, K.F., Tavakkoli, A., Zuckerbrod, S.L.: Attention2angiogan: Synthesizing fluorescein angiography from retinal fundus images using generative adversarial networks. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 9122–9129 (2021). https://doi.org/10.1109/ICPR48806.2021.9412428
13. Kong, L., Lian, C., Huang, D., Hu, Y., Zhou, Q., et al.: Breaking the dilemma of medical image-to-image translation. Advances in Neural Information Processing Systems **34**, 1964–1978 (2021)
14. Kylstra, J.A., Brown, J.C., Jaffe, G.J., Cox, T.A., Gallemore, R., Greven, C.M., Hall, J.G., Eifrig, D.E.: The importance of fluorescein angiography in planning laser treatment of diabetic macular edema. Ophthalmology **106**(11), 2068–2073 (1999)
15. Li, B., Xue, K., Liu, B., Lai, Y.K.: Bbdm: Image-to-image translation with brownian bridge diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition. pp. 1952–1961 (2023)
16. Li, M., Huang, K., Xu, Q., Yang, J., Zhang, Y., Ji, Z., Xie, K., Yuan, S., Liu, Q., Chen, Q.: Octa-500: a retinal dataset for optical coherence tomography angiography study. Medical image analysis **93**, 103092 (2024)
17. Li, M., Cai, W., Liu, R., Weng, Y., Zhao, X., Wang, C., Chen, X., Liu, Z., Pan, C., Li, M., et al.: Ffa-ir: Towards an explainable and reliable medical report generation benchmark. In: Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 2) (2021)
18. Liu, J., Qian, Y., Yang, S., Yan, L., Wang, Y., Gao, M., Liu, L., Xiao, Y., Mo, B., Liu, W.: Pathophysiological correlations between fundus fluorescein angiography and optical coherence tomography results in patients with idiopathic epiretinal membranes. Experimental and Therapeutic Medicine **14**(6), 5785–5792 (2017)
19. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
20. Owen, C.G., Rudnicka, A.R., Nightingale, C.M., Mullen, R., Barman, S.A., Sattar, N., Cook, D.G., Whincup, P.H.: Retinal arteriolar tortuosity and cardiovascular risk factors in a multi-ethnic population study of 10-year-old children; the child heart and health study in england (chase). Arteriosclerosis, thrombosis, and vascular biology **31**(8), 1933–1938 (2011)
21. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: 2011 International conference on computer vision. pp. 2564–2571. Ieee (2011)
22. Staal, J., Abramoff, M., Niemeijer, M., Viergever, M., van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. IEEE Transactions on Medical Imaging **23**(4), 501–509 (2004). https://doi.org/10.1109/TMI.2004.825627
23. Van der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E., Yu, T.: scikit-image: image processing in python. PeerJ **2**, e453 (2014)

24. Wang, H., Xing, Z., Wu, W., Yang, Y., Tang, Q., Zhang, M., Xu, Y., Zhu, L.: Non-invasive to invasive: Enhancing ffa synthesis from cfp with a benchmark dataset and a novel network. In: Proceedings of the 1st International Workshop on Multimedia Computing for Health and Medicine. pp. 7–15 (2024)
25. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
26. Yannuzzi, L.A., Ober, M.D., Slakter, J.S., Spaide, R.F., Fisher, Y.L., Flower, R.W., Rosen, R.: Ophthalmic fundus imaging: today and beyond. American journal of ophthalmology **137**(3), 511–524 (2004)
27. Zhang, J., An, C., Dai, J., Amador, M., Bartsch, D.U., Borooah, S., Freeman, W.R., Nguyen, T.Q.: Joint vessel segmentation and deformable registration on multi-modal retinal images based on style transfer. In: 2019 IEEE International conference on image processing (ICIP). pp. 839–843. IEEE (2019)
28. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
29. Zhou, Y., Wagner, S.K., Chia, M.A., Zhao, A., Xu, M., Struyven, R., Alexander, D.C., Keane, P.A., et al.: Automorph: automated retinal vascular morphology quantification via a deep learning pipeline. Translational vision science & technology **11**(7), 12–12 (2022)