Generation of Structurally Realistic Retinal Fundus Images with Diffusion Models

Sojung Go[1]   Younghoon Ji[2]   Sang Jun Park[1†]   Soochahn Lee[3‡]

[1]Department of Ophthalmology, Seoul National University College of Medicine,
Seoul National University Bundang Hospital, Seongnam, Korea.

[2]VUNO, Inc., Seoul, Korea.

[3]School of Electrical Engineering, Kookmin University, Seoul, Korea.

†sangjunpark@snu.ac.kr   ‡sclee@kookmin.ac.kr

Abstract

We introduce a new technique for generating retinal fundus images that have anatomically accurate vascular structures, using diffusion models. We generate artery/vein masks to create the vascular structure, which we then condition to produce retinal fundus images. The proposed method can generate high-quality images with more realistic vascular structures and can create a diverse range of images based on the strengths of the diffusion model. We present quantitative evaluations that demonstrate the performance improvement using our method for data augmentation on vessel segmentation and artery/vein classification. We also present Turing test results by clinical experts, showing that our generated images are difficult to distinguish with real images. We believe that our method can be applied to construct stand-alone datasets that are irrelevant of patient privacy.

1. Introduction

Retinal fundus images provide clear and high-resolution images of blood vessels in the eye. They are noninvasive and inexpensive. They are used for retinal disease diagnoses and in the early detection and prevention of chronic diseases like diabetes and hypertension.

Extensive research has been conducted on the application of machine learning to fundus images, including tasks such as vessel segmentation [3, 31], artery/vein (A/V) classification [23], and disease classification [19]. The 2D nature of these images, low noise levels, and the seemingly basic structures of the regions of interest, compared to other modalities, may have made them an attractive area of investigation.

Despite the significant interest, limited availability of training data remains a challenge due to the difficulty in constructing the GT. This is particularly true for tasks such as vessel segmentation and A/V classification, which require annotating the vascular structure. Manual or semi-automatic annotation by clinical experts is labor-intensive, making it prohibitively expensive for large-scale datasets. Consequently, most publicly available datasets, including STARE [10], DRIVE [30], CHASEDB1 [7], HRF [2] for vessel segmentation, RITE [11], IOSTAR [34] for A/V classification, and RETA [22] for both tasks, comprise less than one hundred images. Although the recently presented FIVES [13] dataset comprises 800 images, the size is limited compared to datasets such as ImageNet [5] or COCO [21].

Recently, a framework for vessel segmentation and A/V classification based on aligned corresponding fundus and FA images [8, 24, 26, 27] was proposed. The results can serve as an initial baseline, but manual inspection and editing are still required before they can be deemed accurate GT vessel annotations, thereby limiting its scalability. Furthermore, privacy concerns may also limit the dataset size.

To address the need for more training images, generative methods can be used to synthesize additional data. A number of methods have been proposed for this purpose using various GAN models [12, 15]. Early methods, such as those by Zhao et al.[35] and Costa et al.[4], involved a two-step synthesis process comprising vessel mask generation and fundus image generetation. The first step was to synthesize the vessel segmentation mask, using adversarial autoencoders, followed by color fundus image generation using style transfer or conditional GAN. More recent methods, such as those by Yu et al.[33] and Paolo et al.[1], have built on these earlier works, refining the models used to gen-

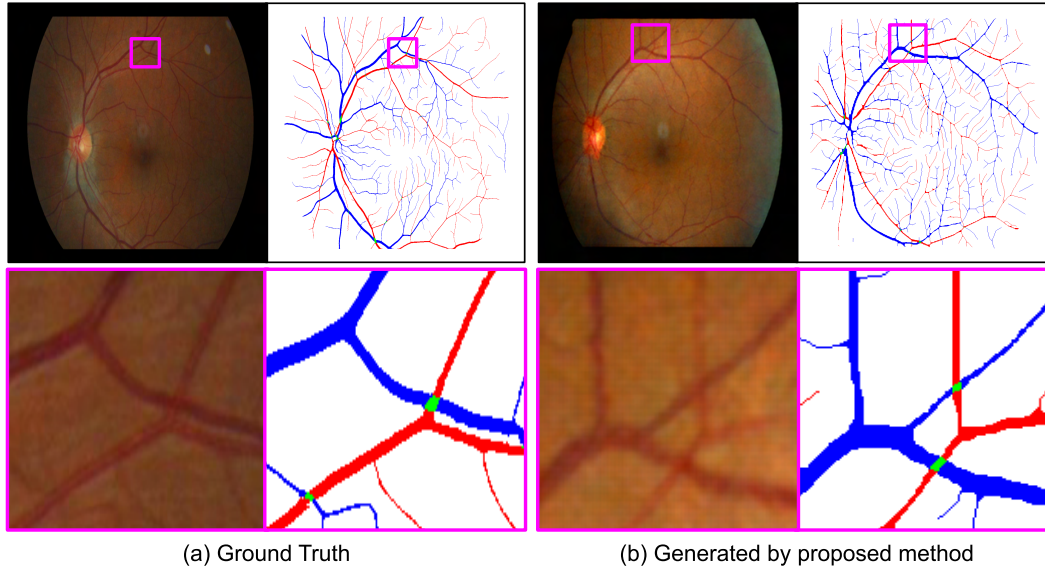(a) Ground Truth   (b) Generated by proposed method

Figure 1. Sample of generated fundus image and artery/vein mask from the proposed method, together with a real data for comparison. The proposed method can generate anatomically accurate fundus images and corresponding artery/vein masks.

erate the vessel mask or fundus image.

Kim et al.[18] proposed a one-step approach, training a StyleGAN[16] on a large dataset of fundus images, that generates retinal images. While their approach generated high-quality images, it lacks the ability to generate GT vessel masks.

Recently, diffusion models [9, 29] have been gaining much attention as an effective generative methodology and an alternative to GANs. Diffusion models often provide better stability in training compared to GANs [6].

Thus, we propose a novel method for generating retinal fundus images with anatomically realistic vascular structures using diffusion models. Our approach consists of two main steps: (1) unconditional generation of A/V masks to establish the vascular structure, (2) generation of retinal fundus images conditioned on the A/V mask with resolution enhancement of the generated image. To address the sparsity of the A/V mask in the first step, we introduce a modified loss function for the diffusion model. Sample results of our method are depicted in Fig. 1.

The main contributions of the proposed method are as follows:

- The proposed method can generate fundus images with more realistic vessel structures since it is trained to generate A/V masks, rather than vessel masks.
- It also generates a wider variety of high-quality images by leveraging the strengths of the diffusion model.
- The generated fundus image can be used directly as GT for vessel segmentation and A/V classification by assigning the generated A/V mask as the GT.

2. Methods

In this section, we provide a detailed description of our proposed method, including the steps involved in collecting the training data. We present a graphical overview of the proposed approach in Figure 2.

2.1. Constructing the training dataset

To implement the proposed method, it is necessary to have access to training datasets consisting of both fundus images and corresponding A/V masks. To generate high-quality A/V masks, we adopt the approach developed by Noh et al. [8, 24, 26, 27], which utilizes fundus images and fluorescein angiography (FA). In the following, we provide a brief overview of this process, which can be divided into two sub-processes for constructing the GT vessel mask [24, 26] and A/V masks [8, 27].

To construct the GT vessel mask, we first construct an aggregated vessel mask from the FA sequence frames.

Rigid registration is performed to register all frames in the FA image sequence into a common reference frame.

Then, vessel segmentation is performed for all frames using a pre-trained CNN. The per-frame alignments are fine-tuned using non-rigid registration on the vessel masks. The aggregate mask is constructed as the maximum vessel probability for each pixel.

The aggregated FA vessel mask is then aligned with the corresponding fundus image using another registration process, and then subjected to simple post-processing for binarization. Because the FA reveals fine vessels more clearly than the fundus images, this mask contains more fine details
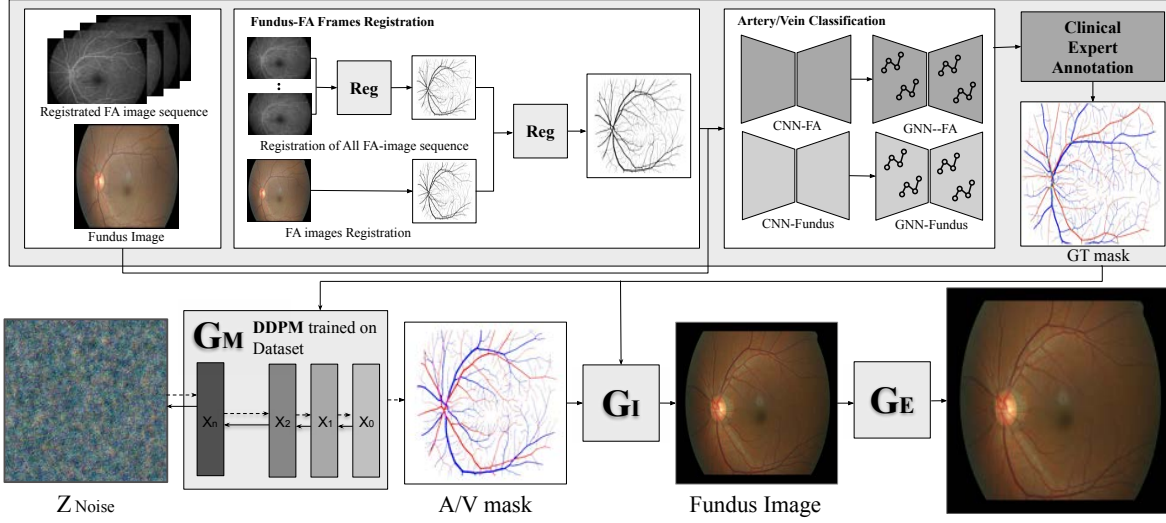
Figure 2. Overview of the proposed method. The top row depicts the process for generating GT A/V masks, which are used to train G_M, which generates new A/V masks, and G_I which generates fundus images conditioned on the A/V masks. G_E performs super-resolution to generate high-resolution images.

than masks generated from fundus images alone. Finally, this baseline vessel mask is manually edited by clinical experts to remove any errors, resulting in the GT vessel mask.

The A/V mask is constructed using the hierarchical vessel graph network (HVGN), which comprises a series of a CNN and a GNN. First, CNN features are constructed for the fundus and FA sequence. The GT vessel mask is used to extract vessel pixels, which are assigned as the vertices of a GNN, with the attributes defined as the CNN features. Edges are assigned between pixels within local window neighborhoods. The CNN and GNN are trained end-to-end with GT A/V masks, so that each node is classified as a label of either artery or vein. We note that each A/V mask is a two-channel binary mask with each channel for the artery mask and vein mask, respectively. This baseline A/V mask is again manually edited by clinical experts to remove any errors, resulting in the final GT A/V mask.

2.2. Generating the vascular structure

We train the model to generate vascular structure, denoted as G_M, on the set of the GT A/V masks. We use the denoising diffusion probabilistic model (DDPM) by [9] and improved by [6].

The concept of diffusion was inspired by the diffusion process in thermodynamics [29]. Given an image $\mathbf{x}_0$, we gradually diffuse its structure by iteratively adding random Gaussian noise. It can be shown that this forward diffusion process can be converted into a single equation to generate the diffused image at any timestep t:

$$x_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \text{with } \epsilon \sim \mathcal{N}(0, I), \quad (1)$$

where $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$ and $\alpha_t := 1 - \beta_t$, where β_t is the variance of the noise at each timestep t.

The idea is to learn the process to reverse this iterative diffusion, i.e., the reverse denoising diffusion process, by training a neural network with parameters θ that generates an estimate of the noise ϵ_θ for a given image $\mathbf{x}_t$. In [9], a U-Net [28] model is used, and is trained with the loss function:

$$L_{simple}(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left[\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t\right)\|^2\right]. \quad (2)$$

In our experiments, we found that if we directly train G_M on the A/V masks using the above configuration, many generated samples contained little to no vascular structures at all. As A/V masks are mostly composed of background, with artery and veins comprising only a small portion of the whole area, even an empty image can be a reasonable sample if all pixels are considered equally. We thus modify the loss function $\mathcal{L}$ as follows:

$$L_{vessel}(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left[\exp(c\mathbf{x}_0)\odot\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t\right)\|^2\right], \quad (3)$$

where $\mathbf{x}_0$ is the given A/V mask, and $\exp(\mathbf{x}_0)$ is the pixel-wise exponential of each pixel value, c is a scalar parameter, and $\odot$ denotes the Hadamard, or element-wise, product. Also, we empirically determined the resolution of the model output and thereby the generated images as 256×256, as the model was difficult to train for higher resolutions. Note that, we incorporate the increased number of attention heads and attention layers, and the adaptive group norm layers as proposed in [6].

2.3. Generating the fundus image

Given the A/V masks, we then train the model to generate the fundus image, denoted as G_I. We apply a conditional GAN, namely, the pix2pixHD [32] as the network model. It is an improvement over the previous pix2pix [12] by using a coarse-to-fine generator, a multi-scale discriminator architecture, and a robust adversarial learning objective function. G_I is trained with GT A/V masks as the input training set and the corresponding fundus images as the output supervision. Since generated A/V masks with low-resolution 256×256 will be given as input at inference time, we resize the GT A/V masks and fundus images to have the same size.

To enhance the low-resolution of the generated images, we further train a super-resolution network, denoted as G_E, to up-sample the images at rate 4 to 1024×1024. We apply the Enhanced Deep Residual Network (EDSR) [20]. The GT fundus images are used as supervision for corresponding input images down-sampled by rate 4.

3. Experiments

3.1. Dataset and implementation details

Private Dataset: We use a private dataset of corresponding fundus images and FA comprising 449 cases of corresponding fundus image and FA pairs from 339 patients from Seoul National University Bundang Hospital (IRB no. B-1810-501-104, approval date 12 Oct, 2018). *. Canon CF60Uvi, Kowa VX-10, and Kowa VX-10a cameras were used to acquire the fundus images and FA. Image resolutions were all normalized to 1536×1024 from original resolutions, varying from 1604×1216 to 2144×1424. We note that we have named this dataset as the FIREFLY (Fundus Images REgistered with FLuorescein angiographY) dataset, and will refer to it as such. For training the CNN for vessel segmentation within the process for constructing the GT vessel segmentation masks of our private SNUBH dataset, we use the publfic datasets DRIVE [30] and HRF [2]. For training the HVGN for constructing the GT A/V masks, we use 46 cases among the 449 of the private dataset.

Public Dataset: We also trained and evaluated our method on the DRIVE public dataset, which includes expert annotated vessel mask GT. Annotated artery/vein vessel mask GT were also made available from added research, entitled as the RITE dataset [11]. The dataset consists of 20 images each for the training and testing datasets.

Environment and implementation details: Experiments were performed on a system with a 2.2 GHz Intel Xeon

---

CPU and 4 nVidia Titan V GPUs. We used the DDPM implementation publicly provided by [6]. We set the maximum number of timestep iterations $T = 1000$ for sampling the DDPM.

3.2. Qualitative Evaluations

Qualitative visual comparisons: We first present qualitative examples of the generated images in Fig. 3. Each column presents three samples generated using a diffusion model, (a) without any structural conditions, (b) conditioned on A/V masks from StyleGAN2, (c) conditioned on vessel masks from a diffusion model, and (d) the proposed method of images conditioned on A/V masks from a diffusion model. For each sample, we present the global image (best viewed using a monitor with by zoom-in) above and a crop-and-zoomed patch below. The zoomed patches for columns (a), (b), and (c) all depict vascular structures with anatomically infeasible structures, including fragmented vessels, unrealistic bifurcations or trifurcations, or unnatural loops. In general, we observed that the generated images from the proposed method generally has more realistic vascular structures.

We also present comparisons between the A/V masks generated from (a) StyleGAN2 and (b) our trained diffusion model in Fig. 4. We can observe that StyleGAN2 mostly fails in generating contiguous vascular structures, especially in the vessel branching, and cannot model crossings. In contrast, A/V masks generated by diffusion have realistic bifurcations and crossings. Also, results from StyleGAN2 are generally more noisy and cannot be readily binarized, compared to images generated by the diffusion.

Qualitative Turing test: We next present results of qualitative Turing tests. We asked five board certified ophthalmologists to determine if the image is from a real patient or was synthesized for (1) real fundus images acquired from patients, (2) fundus images generated from a diffusion model directly trained on fundus images, (3) fundus images generated from A/V masks generated using StyleGAN2 [17], and (4) fundus images generated from A/V masks generated using a diffusion model. For each category, 20 images were presented.

Table 1 summarizes the results. It is evident that there are no decisive image features for identifying real fundus images, as the experts mistook real images as synthetic images in 40% of the cases. Nonetheless, they were able to correctly identify 90% of the images generated from diffusion without generating vascular structure, likely due to unrealistic artifacts. For images generated from the proposed method, the participants were incorrect in more than half the cases. For images generated using StyleGAN2 [17], the expert participants guessed that a higher portion were real compared to the real fundus images. We believe that these results demonstrate that more realistic fundus images are

(a) No Structural Condition   (b) Conditioned on Diffusion Vessel Mask   (c) Conditioned on StyleGAN2 A/V Mask   (d) Conditioned on Diffusion A/V Mask
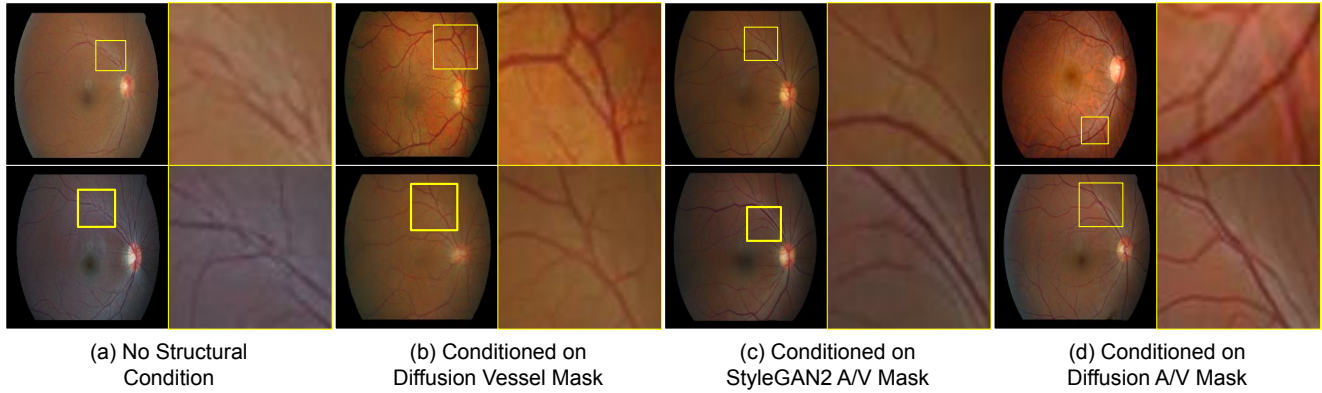
Figure 3. Samples of generated fundus images for qualitative evaluation. Two samples of generated fundus images (a) with no structural condition, (b) conditioned on vessel masks generated by diffusion model, (c) conditioned on A/V masks generated by StyleGAN2, and, (d-proposed) conditioned on A/V mask generated by diffusion model.

(a) StyleGAN2
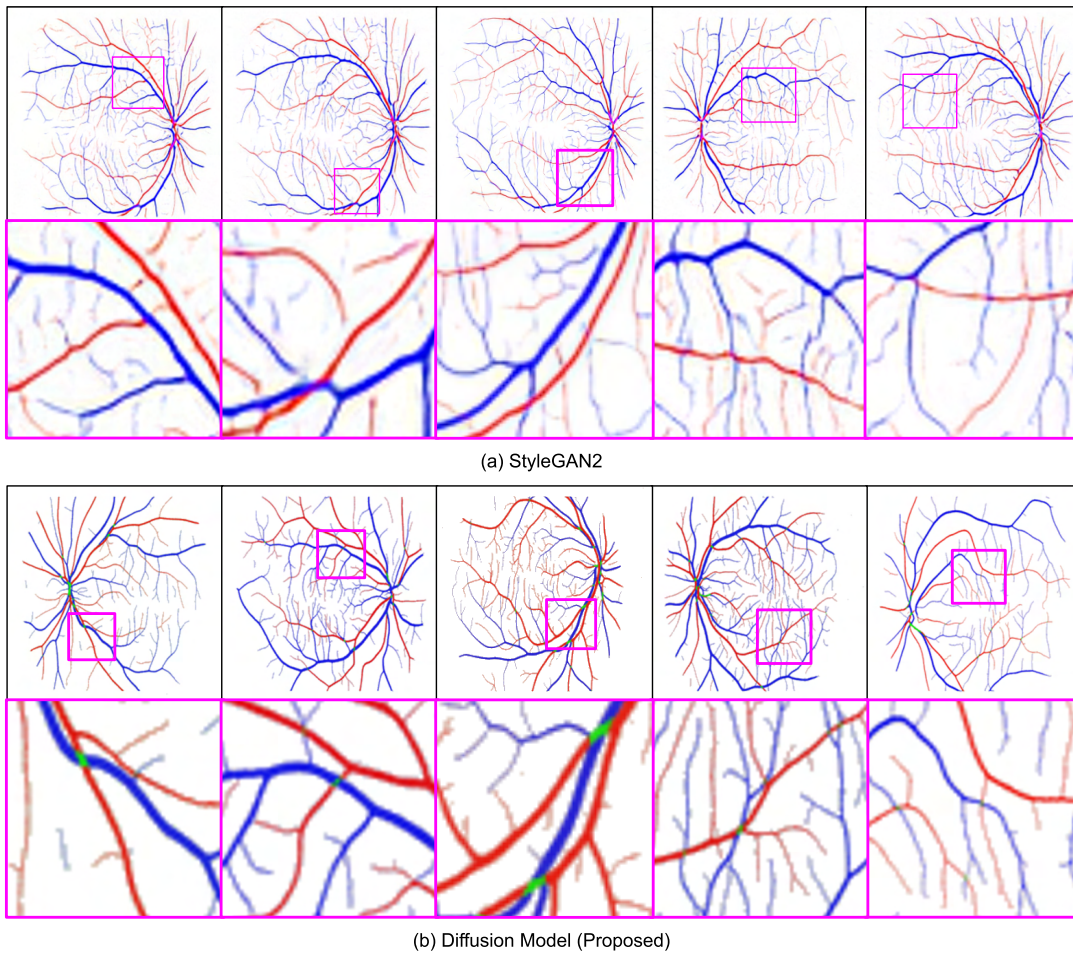

(b) Diffusion Model (Proposed)

Figure 4. Samples of generated artery/vein masks from StyleGAN2 and the diffusion model. Masks from the diffusion model have fewer discontinuities and include realistic vessel crossings. Hyperparameters for both methods have been empirically optimized.

generated when conditioned on vascular structures.

Table 1. Qualitative comparative evaluation of generated fundus images via Turing test, asking whether the image is real or synthesized. Each test comprises 20 images.

Real/Synth.	Gen. Model	Vascular structure	Acc.(95% CI)
Real	-	-	60($\pm$11.47)
Synth	Diffusion [6]	None	90($\pm$8.99)
Synth	Diffusion [6]	A/V mask	41.25($\pm$7.35)
Synth	StyleGAN2 [17]	A/V mask	22.5($\pm$9.34)

3.3. Quantitative Evaluations

As it was observed in the qualitative Turing test, it is very difficult to estimate or measure how realistic is the generated image. We thus focus on measuring the effect of the generated images on the original goal, of improving the training of methods of vessel segmentation and A/V classification.

In Table 2, we present the performance of recent state-of-the-art methods SSA-Net [25] and RV-GAN [14] for vessel segmentation, and HVGN [8] for A/V classification, with and without data augmentation based on the generated fundus images and corresponding vessel masks. For all methods, we use the implementations provided by the authors. For the HVGN method, we evaluate the version with only fundus image inputs, in contrast to the version that with both fundus and FA images, which was used to construct the GT A/V masks of the FIREFLY dataset.

We present results using the private FIREFLY dataset and the public DRIVE [30]/RITE [11] dataset. For the FIREFLY dataset, we assign a random partition of 100 images for the test set and the remaining 359 images for the training set. We present results for within-domain evaluation, i.e., using the training and test sets from the same dataset, as well as the cross-domain evaluation trained on the FIREFLY training set and tested on the DRIVE/RITE test set. For all configurations, 3000 synthetic A/V masks and images and were generated from the proposed method were used in the data augmentation. Accuracy (Acc), sensitivity (Se), specificity (Sp), and the area-under-curve (AUC) of the receiver-operating-characteristic (ROC) curve are measured based on pixel-wise true/false positives and true/false negatives, for binary vessel masks and A/V masks comprising binary artery and vein tree masks, respectively.

We observed that the quantitative results were improved for all measurements when using the generated data augmentation, regardless of the vessel analysis task or specific method, or the specific train/test dataset configuration. We believe these results strongly support the effectiveness and generalizability of the proposed method.

In Table 3, we present the comparative evaluation according to the number of the synthesized images for data augmentation, for the FIREFLY dataset. Similar to a practical dataset expansion process, we simulate a gradual increase of the dataset, adding 1000 generated images up to a maximum of 6000 images. We train methods for vessel segmentation and A/V classification, SSA-Net [25] and HVGN [8], respectively, on these datasets. We observed that the quantitative results were improved for almost all measurements with the increase in the synthetic data. Again, we believe that these results support the relevance of the generated images in training methods for vessel analysis.

To further evaluate how much the generated images can faithfully constitute relevant information, we trained the vessel segmentation and A/V classification methods on synthetic images only, without any real images. We observed that there were only slight decreases in the quantitative results, demonstrating that the generated images faithfully represent the characteristics of the original real images, and contain sufficient information for training methods for vessel analysis.

We provide sample results for vessel segmentation and A/V classification in Fig. 5 and Fig. 6, respectively. These results demonstrate that the data augmentation from the images generated from the diffusion model helps to correct many misclassifications and broken vessel branches in vessel segmentation and A/V classification.

Quantitative evaluation of perceptual image quality: In Table 4, we present the FID values for the fundus images generated from A/V masks generated using Style-GAN2 [17], and those generated from A/V masks generated using diffusion [6]. It is evident that the diffusion model results in lower distances, demonstrating the benefit of the diffusion model. We note that, due to computational constraints, neither GAN nor diffusion model were not optimized in respect to the scores, as was done in [18].

4. Discussion

Different models may be used as the generators for the mask, G_M, the mask-to-image generator G_I, and the high-res image generator G_E. For instance, diffusion models could also be used for both G_I and G_E. Here, we avoided their use, because they required considerably higher computation, while only achieving comparable results. This may be due to the nature of the GT fundus images, which were limited to images from normal patients, and mostly contains

Table 2. Comparative evaluation of vessel analysis methods trained w/wo generative data augmentation using on the proposed method. Area-under-curve (AUC) of the receiver-operation-characteristic curve (ROC), Accuracy (Acc), Sensitivity (Se), and specificity (Sp) are presented.

Task	Model	Train DB*	Aug.	Test DB	AUC	Acc.	Se.	Sp.
Segmentation	SSA-Net [25]	FIREFLY	✗	FIREFLY	0.973	0.963	0.797	0.983
			✓		0.984+.011	0.976	0.815	0.990
		FIREFLY	✗	DRIVE-AV	0.948	0.910	0.870	0.920
			✓		0.952+.004	0.918	0.874	0.924
		DRIVE-AV	✗	DRIVE-AV	0.961	0.914	0.868	0.921
			✓		0.966+.005	0.921	0.906	0.923
	RV-GAN [14]	FIREFLY	✗	FIREFLY	0.937	0.945	0.713	0.989
			✓		0.944+.007	0.947	0.672	0.993
		DRIVE-AV	✗	DRIVE-AV	0.970	0.950	0.630	0.919
			✓		0.981+.011	0.967	0.658	0.995
A/V cls.	HVGN [8]	FIREFLY	✗	FIREFLY	0.973	0.912	0.923	0.901
			✓		0.980+.007	0.922	0.930	0.915
		FIREFLY	✗	DRIVE-AV	0.909	0.819	0.971	0.768
			✓		0.914+.005	0.824	0.879	0.769
		DRIVE-AV	✗	DRIVE-AV	0.917	0.855	0.877	0.832
			✓		0.924+.007	0.860	0.884	0.835

*The diffusion model for image generation and the model for vessel analysis tasks are both trained using the same *Train DB*. A total of 3000 synthetic A/V masks and fundus images generated by the proposed method were used for data augmentation.

Table 3. Comparative evaluation of vessel analysis methods on the FIREFLY dataset, trained with different number of generated images in data augmentation. Area-under-curve (AUC) of the receiver-operation-characteristic curve (ROC), Accuracy (Acc), Sensitivity (Se), and specificity (Sp) are presented.

Training Set Size	Segmentation SSA-Net [25]				A/V cls. HVGN [8]			
	AUC	Acc.	Se.	Sp.	AUC	Acc.	Se.	Sp.
Real. 359	0.973	0.963	0.797	0.983	0.973	0.912	0.923	0.901
Real + Synth. 500	0.979	0.972	0.798	0.986	0.976	0.915	0.926	0.905
Real + Synth. 1000	0.980	0.973	0.803	0.988	0.976	0.915	0.927	0.902
Real + Synth. 2000	0.983	0.975	0.812	0.989	0.978	0.921	0.928	0.914
Real + Synth. 3000	0.984	0.976	0.815	0.990	0.980	0.922	0.930	0.915
Real + Synth. 6000	**0.986**	**0.978**	**0.819**	**0.991**	**0.984**	0.930	**0.935**	**0.926**
Synth. 6000	0.985	0.976	0.818	0.990	0.983	**0.932**	0.933	0.924
Synth. 3000	0.984	0.975	0.816	0.988	0.980	0.921	0.929	0.912
Synth. 2000	0.982	0.974	0.811	0.989	0.978	0.919	0.926	0.910
Synth. 1000	0.979	0.973	0.801	0.988	0.975	0.914	0.924	0.903

Table 4. Fréchet Inception Distance (FID) for generated fundus images.

Gen. Model	Vascular structure	FID
StyleGAN2 [17]	A/V mask	122.8
Diffusion [6]	A/V mask	86.78

smooth textures other than the vascular structure. As our main target application was the use of generated images for training methods for vessel analysis, we did not consider various pathologies such as macular degeneration. Thus, further research may be warranted for improving the generation framework when generating synthetic datasets for disease diagnosis.

In the short term, we plan to apply the proposed method to construct a large-scale dataset with high-quality GT vascular structures and without any privacy concerns. Based on the dataset, we hope to further develop more advanced metrics for measuring the performance of segmentation and classification techniques.
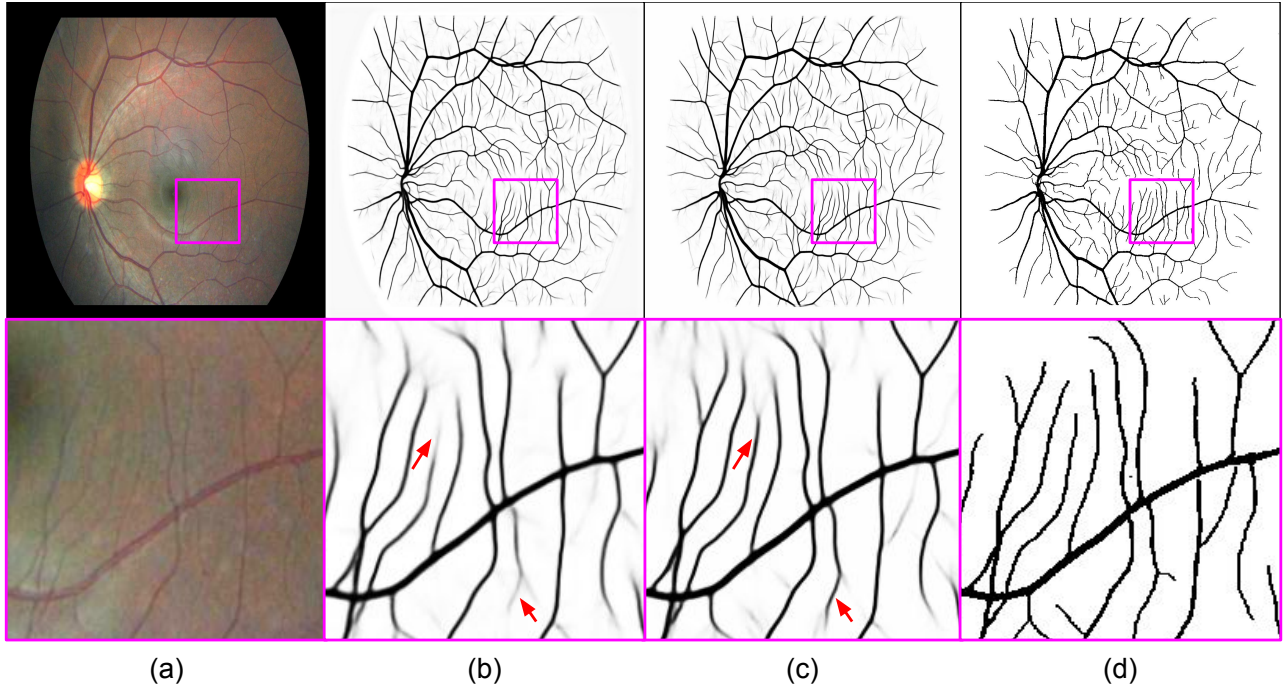
Figure 5. Sample vessel segmentation result of SSA-Net [25], trained on the FIREFLY dataset. (a) Input fundus image, (b) result of SSA-Net trained only on real images, (c) result of SSA-Net trained with data augmentation from synthetic dataset of 3000 images generated by the proposed method, and (d) the GT. The top and bottom rows show the images in full and zoomed resolution.
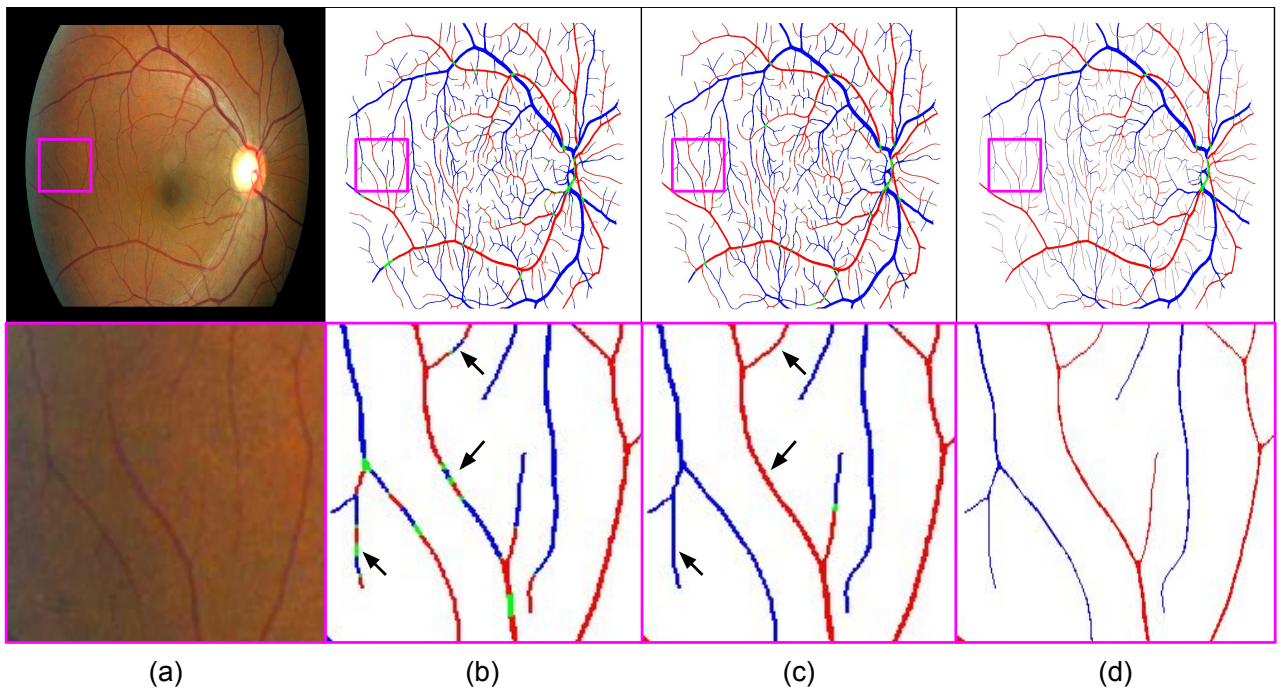

Figure 6. Sample A/V classification result of HVGN [8], trained on the FIREFLY dataset. (a) Input fundus image, (b) result of HVGN trained only on real images, (c) result of HVGN trained with data augmentation from synthetic dataset of 3000 images generated by the proposed method, and (d) the GT. The top and bottom rows show the images in full and zoomed resolution.

References

[1] Paolo Andreini, Giorgio Ciano, Simone Bonechi, Caterina Graziani, Veronica Lachi, Alessandro Mecocci, Andrea Sodi, Franco Scarselli, and Monica Bianchini. A two-stage gan for high-resolution retinal image generation and segmentation. *Electronics*, 11(1), 2022. 1

[2] A. Budai, R. Bock, A. Maier, J. Hornegger, and G. Michelson. Robust vessel segmentation in fundus images. *Int. J. Biomed. Imaging*, 2013:154860, 2013. 1, 4

[3] Chunhui Chen, Joon Huang Chuah, Raza Ali, and Yizhou Wang. Retinal vessel segmentation using deep learning: A review. *IEEE Access*, 9:111985–112004, 2021. 1

[4] Pedro Costa, Adrian Galdran, Maria Ines Meyer, Meindert Niemeijer, Michael Abràmoff, Ana Maria Mendonça, and Aurélio Campilho. End-to-end adversarial retinal image synthesis. *IEEE Trans. Med. Imaging*, 37(3):781–791, 2018. 1

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*, pages 248–255, 2009. 1

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, pages 8780–8794, 2021. 2, 3, 4, 6, 7

[7] M. M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A. R. Rudnicka, C. G. Owen, and S. A. Barman. An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE Trans. Biomed. Eng.*, 59(9):2538–2548, 2012. 1

[8] Sojung Go, Jooyoung Kim, Kyoung Jin Noh, Sang Jun Park, and Soochahn Lee. Combined deep learning of fundus images and fluorescein angiography for retinal artery/vein classification. *IEEE Access*, 10:70688–70698, 2022. 1, 2, 6, 7, 8

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 2, 3

[10] A. D. Hoover, V. Kouznetsova, and M. Goldbaum. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Trans. Med. Imaging*, 19(3):203–210, 2000. 1

[11] Qiao Hu, Michael D. Abràmoff, and Mona K. Garvin. Automated separation of binary overlapping trees in low-contrast color retinal images. In *MICCAI*, pages 436–443, 2013. 1, 4, 6

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE CVPR*, pages 5967–5976, 2017. 1, 4

[13] Kai Jin, Xingru Huang, Jingxing Zhou, Yunxiang Li, Yan Yan, Yibao Sun, Qianni Zhang, Yaqi Wang, and Juan Ye. Fives: A fundus image dataset for artificial intelligence based vessel segmentation. *Scientific Data*, 9(1):475, 2022. 1

[14] Sharif Amit Kamran, Khondker Fariha Hossain, Alireza Tavakkoli, Stewart Lee Zuckerbrod, Kenton M Sanders, and Salah A Baker. Rv-gan: Segmenting retinal vascular structure in fundus photographs using a novel multi-scale generative adversarial network. In *MICCAI*, pages 34–44. Springer, 2021. 6, 7

[15] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018. 1

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF CVPR*, pages 4396–4405, 2019. 2

[17] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020. 4, 6, 7

[18] Mingyu Kim, You Na Kim, Miso Jang, Jeongeun Hwang, Hong-Kyu Kim, Sang Chul Yoon, Yoon Jeon Kim, and Namkug Kim. Synthesizing realistic high-resolution retina image by style-based generative adversarial network and its utilization. *Scientific Reports*, 12(1):17307, 2022. 2, 6

[19] Tao Li, Wang Bo, Chunyu Hu, Hong Kang, Hanruo Liu, Kai Wang, and Huazhu Fu. Applications of deep learning in fundus images: A review. *Med. Image Anal.*, 69:101971, 2021. 1

[20] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *IEEE CVPR Workshops*, pages 1132–1140, 2017. 4

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014. 1

[22] Xingzheng Lyu, Li Cheng, and Sanyuan Zhang. The reta benchmark for retinal vascular tree analysis. *Scientific Data*, 9(1):397, 2022. 1

[23] Muthu Rama Krishnan Mookiah, Stephen Hogg, Tom J MacGillivray, Vijayaraghavan Prathiba, Rajendra Pradeepa, Viswanathan Mohan, Ranjit Mohan Anjana, Alexander S. Doney, Colin N.A. Palmer, and Emanuele Trucco. A review of machine learning methods for retinal blood vessel segmentation and artery/vein classification. *Med. Image Anal.*, 68:101905, 2021. 1

[24] Kyoung Jin Noh, Sang Jun Park, and Soochahn Lee. Fine-scale vessel extraction in fundus images by registration with fluorescein angiography. In *MICCAI*, pages 779–787, 2019. 1, 2

[25] Kyoung Jin Noh, Sang Jun Park, and Soochahn Lee. Scale-space approximated convolutional neural networks for retinal vessel segmentation. *Comput. Methods Programs Biomed.*, 178:237–246, 2019. 6, 7, 8

[26] Kyoung Jin Noh, Jooyoung Kim, Sang Jun Park, and Soochahn Lee. Multimodal registration of fundus images with fluorescein angiography for fine-scale vessel segmentation. *IEEE Access*, 8:63757–63769, 2020. 1, 2

[27] Kyoung Jin Noh, Sang Jun Park, and Soochahn Lee. Combining fundus images and fluorescein angiography for artery/vein classification using the hierarchical vessel graph network. In *MICCAI*, pages 595–605, 2020. 1, 2

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 3

[29] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using

nonequilibrium thermodynamics. In *ICML*, pages 2256–2265, Lille, France, 2015. PMLR. 2, 3

[30] J. Staal, M. D. Abramoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE Trans. Med. Imaging*, 23 (4):501–509, 2004. 1, 4, 6

[31] Olubunmi Omobola Sule. A survey of deep learning for retinal blood vessel segmentation methods: Taxonomy, trends, challenges and future directions. *IEEE Access*, 10:38202–38236, 2022. 1

[32] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE/CVF CVPR*, pages 8798–8807, 2018. 4

[33] Zekuan Yu, Qing Xiang, Jiahao Meng, Caixia Kou, Qiushi Ren, and Yanye Lu. Retinal image synthesis from multiple-landmarks input with generative adversarial networks. *Biomed. Eng. OnLine*, 18(1):62, 2019. 1

[34] J. Zhang, B. Dashtbozorg, E. Bekkers, J. P. W. Pluim, R. Duits, and B. M. ter Haar Romeny. Robust retinal vessel segmentation via locally adaptive derivative frames in orientation scores. *IEEE Trans. Med. Imaging*, 35(12):2631–2644, 2016. 1

[35] He Zhao, Huiqi Li, Sebastian Maurer-Stroh, and Li Cheng. Synthesizing retinal and neuronal images with generative adversarial nets. *Med. Image Anal.*, 49:14–26, 2018. 1