DiffDGSS: Generalizable Retinal Image Segmentation with Deterministic Representation from Diffusion Models

Yingpeng Xie[1], Junlong Qu[1], Hai Xie[1], Tianfu Wang[1*], and Baiying Lei[1*]

National-Regional Key Technology Engineering Laboratory for MedicalUltrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Shenzhen University Medical school, Shenzhen 518060, China
tfwang@szu.edu.cn; leiby@szu.edu.cn

Abstract. Acquiring a comprehensive segmentation map of the retinal image serves as the preliminary step in developing an interpretable diagnostic tool for retinopathy. However, the inherent complexity of retinal anatomical structures and lesions, along with data heterogeneity and annotations scarcity, poses challenges to the development of accurate and generalizable models. Denoising diffusion probabilistic models (DDPM) have recently shown promise in various medical image applications. In this paper, driven by the motivation to leverage strong pre-trained DDPM, we introduce a novel framework, named DiffDGSS, to exploit the latent representations from the diffusion models for Domain Generalizable Semantic Segmentation (DGSS). In particular, we demonstrate that the deterministic inversion of diffusion models yields robust representations that allow for strong out-of-domain generalization. Subsequently, we develop an adaptive semantic feature interpreter for projecting these representations into an accurate segmentation map. Extensive experiments across various tasks (retinal lesion and vessel segmentation) and settings (cross-domain and cross-modality) demonstrate the superiority of our DiffDGSS over state-of-the-art methods.

Keywords: Retinal Image · Diffusion Models · Latent Representations · Domain Generalizable Semantic Segmentation.

1 Introduction

Retinal fundus images give access to a highly detailed view of the interior surface of the eye, typically centered around the macula or optic disc, and contain several diagnostically relevant biomarkers. An abnormality in the retina can either be a manifestation of eye disease, systemic disease, or trauma-induced injuries [22]. Therefore, accurately segmenting fundus images is a foundational step toward creating an interpretable diagnostic tool. However, the intricate nature of retinal anatomical structures and lesions, combined with the variability across datasets,

as well as the scarcity of annotations, present significant obstacles in developing models that are both accurate and capable of generalizing well.

With the intuition that the ability to generate data from a given domain implies a profound understanding of the semantics of that domain, a line of works has investigated the intermediate representations derived from Generative Adversarial Networks (GANs) [7]. It was shown that these representations can be decoded to produce a semantic segmentation map of the image, and training the decoder requires only a handful of labeled examples to generalize to the rest of the latent representations [30]. However, due to its lack of inference functionality, extracting the latent representations for real samples necessitates the use of GANs inversion [27], which inverts the sample back into the latent space of trained GANs. Current GANs inversion techniques either suffer from limited reconstruction quality or require significantly higher computational costs [27]. In recent years, Denoising Diffusion Probabilistic Models (DDPM) [8] have been introduced as a more effective form of generative modeling, which learns a network that iteratively predicts and removes noise of multiple levels driven by a diffusion process. In this context, Baranchuk *et al.* [1] investigated the intermediate representations within pre-trained diffusion models that perform the Markov step of the reverse diffusion (generative) process. These representations have been found to be useful beyond noise prediction and possess greater semantic significance than those of GANs, allowing the segmentation branch to produce very accurate labels for real images and not suffer from error-prone GANs inversion. However, this generative process is inherently stochastic, where the latent space only consists of a sequence of noise-induced degradation with limited semantic content and can not be used to reconstruct observations [24].

Comprehending the latent space of diffusion models is crucial but challenging, and it is the key to advancing the use of diffusion models. By generalizing the forward diffusion process from a Markov chain to a Non-Markov one, Song *et al.* [24] showed that every stochastic DDPM has an ODE-based, deterministic counterpart with the same output distribution. An important advantage of the ODE-based, deterministic DDPM is that the generative process can be inverted. Specifically, by using the deterministic inversion technique derived from Denoising Diffusion Implicit Models (DDIM), one can retrieve the latent code of a given image, which can then be denoised to reconstruct the image [24]. More importantly, unlike GANs the sophisticated inversion method is required, it can retrieve the latent code of an arbitrary real image even if the image is not in the trained domain [13]. Motivated by this insight, in this paper, we delve into the intermediate representations derived from this process, with a particular focus on Domain-Generalizable Semantic Segmentation (DGSS).

One major problem is that diffusion models learn visual concepts by solving pretext tasks at thousands of noise levels, and it is not clear what information the model learns at each level during training, and thus hard to determine the priority of each timestep. Besides, the existing practice of upsampling feature maps from various stages of the decoder to the highest resolution, followed by a straightforward concatenation of these maps across multiple timesteps and

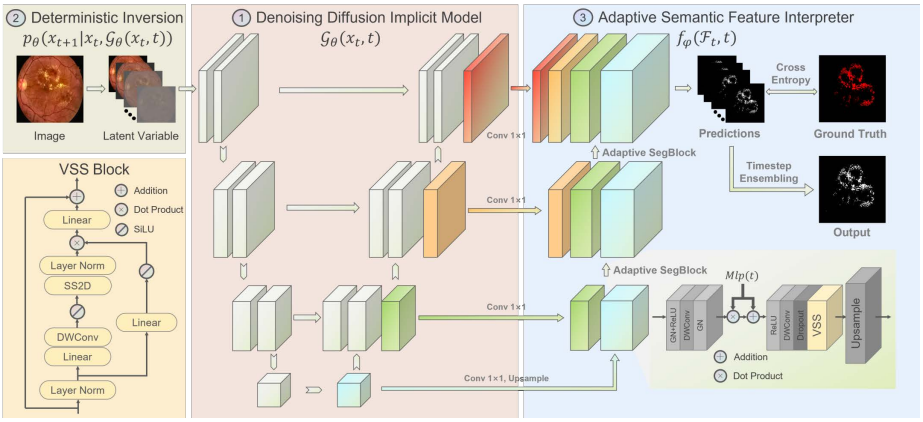

Fig. 1. An overview of our DiffDGSS. Given an off-the-shelf diffusion model, the robust latent representation of the image is obtained from the network by performing deterministic DDIM inversion. Subsequently, we train a feature interpreter branch on top of this multi-scale and timestep-dependent representation to predict a segmentation map.

blocks, leads to significant computer memory usage [1]. Also, they usually employ an ensemble of multi-layer perceptions (MLPs) to independently interpret the feature vector of each pixel into a pixel-wise label [30, 1], which fails to integrate local features with their global dependencies that are crucial for accurately segmenting medical images. To tackle these challenges, we develop a simple but effective interpreter to precisely refine and forward the feature maps from each layer to the next layer within the DDPM backbone, ultimately yielding the segmentation map. Overall, the main contributions of this paper can be summarized as follows: (1) We present DiffDGSS, an innovative representation-based approach aimed at achieving precise retinal image segmentation and robust out-of-domain generalization; (2) We design an adaptive semantic feature interpreter to decode the multi-scale representations and dynamically adapt its behavior over sampling timesteps; (3) Qualitative and quantitative experiments across various tasks and settings demonstrate the superiority of our DiffDGSS over state-of-the-art methods.

2 Methodology

We illustrate the overview of DiffDGSS in Fig. 1. First, we provide a brief overview of the DDPM framework. Subsequently, we delineate the deterministic inversion technique and elaborate on the extraction of deterministic representations from the DDPM backbone. Finally, we present our adaptive feature interpreter, designed to effectively utilize these representations for DGSS.

2.1 Denoising Diffusion Probabilistic Models

Consider a real data point $x_0 \sim q(x_0)$, DDPM delineates a forward diffusion process that sequentially introduces Gaussian noise to the sample over T steps, resulting in a sequence of increasingly noisy samples $x_1, \ldots, x_T$:

$$q\left(x_t \mid x_{t-1}\right) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t \mathbf{I}\right) \tag{1}$$

where $\{\beta_t \in (0,1)\}_{t=1}^{T}$ denotes a fixed variance schedule and T is the maximum timestep chosen so that x_T resembles pure noise. Once the forward diffusion process can be reversed, it becomes possible to create a new sample given a pure noise. However, the reverse diffusion process $q\left(x_{t-1} \mid x_t\right)$ cannot be computed directly, DDPM approximates it via:

$$p_\theta\left(x_{t-1} \mid x_t\right) = \mathcal{N}\left(x_{t-1}; \mu_\theta\left(x_t, t\right), \Sigma_\theta\left(x_t, t\right)\right) \tag{2}$$

where $\mu_\theta\left(x_t, t\right)$ and $\Sigma_\theta\left(x_t, t\right)$ refer to the mean predictor and covariance predictor, respectively. In practice, Ho *et al.* [8] proposed to train a noise predictor $\epsilon_\theta\left(x_t, t\right)$ instead of directly learning $\mu_\theta\left(x_t, t\right)$ and fix the variance $\Sigma_\theta\left(x_t, t\right)$ to a constant since it produces better samples.

2.2 Deterministic Representation from Diffusion Models

By generalizing DDPM via a class of non-Markovian diffusion processes that lead to the same training objective, Song *et al.* [24] proposed a more efficient class of iterative implicit probabilistic models named DDIM that enjoys the following deterministic posterior distribution:

$$q\left(x_{t-1} \mid x_t, x_0\right) = \mathcal{N}\left(\sqrt{\bar{\alpha}_{t-1}} x_0 + \sqrt{1-\bar{\alpha}_{t-1}} \frac{x_t - \sqrt{\bar{\alpha}_t} x_0}{\sqrt{1-\bar{\alpha}_t}}, 0\right) \tag{3}$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, the DDIM sampling (generative) process, anchored by the initial noise x_T, forges an implicit latent space, which corresponds to executing an ODE integration in the forward direction:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \left(\frac{x_t - \sqrt{1-\bar{\alpha}_t} \epsilon_\theta\left(x_t, t\right)}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}} \epsilon_\theta\left(x_t, t\right) \tag{4}$$

Importantly, one can run this generative process in reverse (known as DDIM Deterministic Inversion [24]) to retrieve the latent code capable of reconstructing the real image with high fidelity. Although a slight error is introduced in each step, it works well in the case without classifier-free guidance [20]. Our key insight is that there is a deterministic mapping between the latent space and the image space, and thus we parameterize $\mu_\theta\left(x_t, t\right)$ with an image generator $\mathcal{G}_\theta\left(x_t, t\right)$ [21], which map a series of latent variables to a particular image:

$$\begin{aligned} \mathcal{G}_\theta\left(x_t, t\right) &= \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta\left(x_t, t\right)\right) \\ \mu_\theta\left(x_t, t\right) &= \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathcal{G}_\theta\left(x_t, t\right) + \frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t}x_t \end{aligned} \tag{5}$$

In other words, the network predicts a clean input x_0 given x_t, which generates a diverse set of latent representations in the form of intermediate feature maps [1] while executing the DDIM deterministic inversion:

$$x_{t+1} = \sqrt{\bar{\alpha}_{t+1}}\mathcal{G}_\theta\left(x_t, t\right) + \sqrt{1-\bar{\alpha}_{t+1}}\frac{x_t - \sqrt{\bar{\alpha}_t}\mathcal{G}_\theta\left(x_t, t\right)}{\sqrt{1-\bar{\alpha}_t}} \tag{6}$$

Such timestep-dependent representation allows treating them as deterministic representations of x_0 for domain generalizable semantic segmentation.

2.3 Adaptive Semantic Feature Interpreter

We develop an adaptive semantic feature interpreter $f_\varphi\left(\mathcal{F}_t, t\right)$ to interpret the multi-scale representations $\mathcal{F}_t$ and dynamically adapt its behavior over sampling timesteps t. It conditions the timestep-dependent representations using adaptive group normalization layers (AdaGN), following Dhariwal *et al.* [5], which extend group normalization [26] by applying channel-wise scaling and shifting on the normalized feature maps $\mathbf{h} \in \mathbf{R}^{c\times h\times w}$:

$$\text{AdaGN}\left(\mathbf{h}, t\right) = \mathbf{t}_s\,\text{GroupNorm}(\mathbf{h}) + \mathbf{t}_b \tag{7}$$

where $(\mathbf{t}_s, \mathbf{t}_b) \in \mathbf{R}^{2\times c} = \text{MLP}(\psi(t))$ is the output of a multilayer perceptron with a sinusoidal encoding function ψ. Then, the adaptive segblock comprises two successive DWConv blocks [10] followed by the VSS-based Mamba block [17] for short- and long-range dependency modeling. At inference time, we use a majority voting mechanism to ensemble the prediction map of each timesteps-dependent representation to obtain the final segmentation map [1].

3 Experiments and Results

Dataset Our approach began by pre-training DDPM using the large-scale unlabeled EyePACS dataset ($88,702$ images) [4]. Then, we evaluate our approach on two distinct segmentation tasks of retinal fundus images: lesion segmentation (IDRID Lesion Segmentation Set [23]) and vessel segmentation (including STARE [9], HRF [2], DRIVE [25] and CHASEDB1 [6] Dataset). To evaluate the cross-modality vessel segmentation performance, we include two OCTA datasets for evaluation, namely the OCTA-500 [15] and ROSE [19]. The partitioning of the training and test sets adheres to prior studies [18].

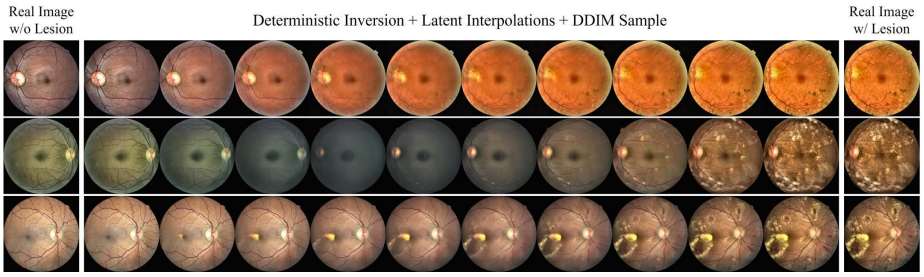

Fig. 2. Semantic interpolation between two real images. We employ the DDIM deterministic inversion technique to obtain the latent codes of given images and interpolate them linearly from one to another, and then we decode them to the image space.

Metrics We adopt the area-under-the-curve (AUC) of both the precision-recall (PR) curve and receiving operating characteristic (ROC) curve to assess the lesion segmentation performance, and employ the Dice Similarity Coefficient (DSC) to assess the vessel segmentation performance, as they are also recognized as metrics in prior competitions and research [31, 18].

Table 1. Comparison results of state-of-the-art methods on the IDRiD dataset. *Method uses in-domain IDRiD Grading Set [23] while †Method uses ImageNet dataset as the unlabeled data. Top 1 results are highlighted in bold.

Method	Soft Exudates		Hemorrhage		Microaneurysms		Hard Exudates	
	ROC	PR	ROC	PR	ROC	PR	ROC	PR
VRT(1st Team)[23]	-	69.95	-	68.04	-	49.51	-	71.27
19'AdvSeg* [12]	93.18	67.56	92.56	59.23	96.12	47.06	94.56	80.32
19'ASDNet* [12]	94.89	69.24	93.24	62.85	96.92	47.82	95.02	80.95
19'Zhou *et al.** [31]	99.36	74.07	97.79	**69.36**	98.28	49.60	99.35	**88.72**
20'Self-training* [32]	-	73.96	-	65.66	-	49.57	-	86.08
22'DDPM-Seg [1]	98.59	76.54	96.54	61.39	97.46	32.36	99.09	82.78
23'FEDD† [3]	97.94	59.64	95.89	52.11	96.34	41.41	99.30	85.04
23'Cut-Paste* [28]	-	76.91	-	66.67	-	**50.20**	-	87.24
DiffDGSS (Ours)	**99.43**	**79.71**	**98.04**	65.44	**99.20**	43.15	**99.52**	88.26
w/o Deterministic Inversion	98.65	77.69	97.53	63.56	97.69	41.00	99.44	86.47
w/o Adaptive Interpreter	99.06	74.38	97.54	62.30	98.41	33.60	99.14	84.30

Implementation Details Preprocessing involves the removal of black boundary regions from retinal images, followed by resizing to 512×512 pixels. Also, we process Contrast Limited Adaptive Histogram Equalization (CLAHE) on all images to enhance image contrast while preserving local details [11]. And we

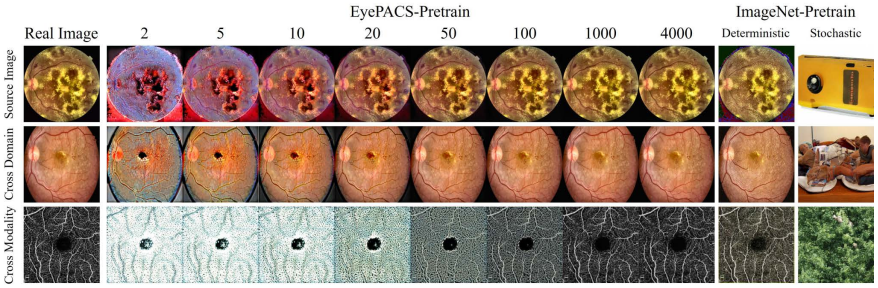

Fig. 3. Visualization of out-of-domain generalization in retinal image reconstruction: A comprehensive analysis of DDIM deterministic inversion and sampling across an extensive range of timesteps and data settings.

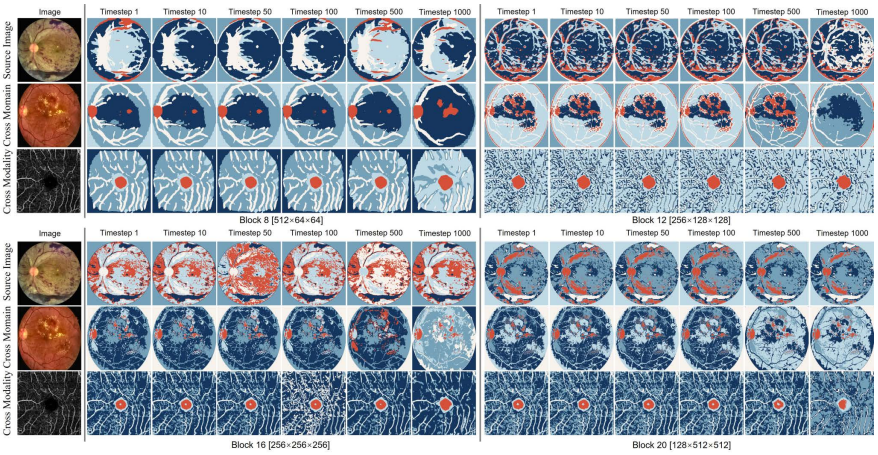

Fig. 4. Visualization of k-means clusters (k=5) formed by the representations at the DDPM decoder blocks $\{8, 12, 16, 20\}$ across diffusion steps $\{1, 10, 50, 100, 500, 1000\}$.

use only flipping for data augmentation [5]. Hyperparameters and model architecture for the pre-training DDPM follow the Guided-diffusion implementation for ImageNet-512 [5] with total iterations of 120K and diffusion steps of 4000. For the training of feature interpreter, through empirical experiments but not tuned for each dataset, the representations are extracted from the DDPM decoder behind the 10th block with timesteps $\{1, 10, 100, 500\}$, and we use the Adam optimizer with a learning rate of 0.0002 and a batch size of 4.

3.1 Experimental results and analysis

Deterministic Representation for DGSS In Fig. 2, the left and rightmost images are real fundus images, and between them are reconstructed interpolations in DDIM latent space. The result demonstrates the remarkable capability

of deterministic inversion to capture and encode intricate high-level semantics inherent in the samples. Furthermore, as shown in Fig. 3, the result underscores that deterministic inversion gives latent that allows for reasonable reconstructions even with a few steps. Importantly, the findings also showcase the latent space's ability to generalize beyond the training domain, successfully preserving and reconstructing the semantic intricacies of retinal images from domains or modalities previously unencountered. Interestingly, even with the ImageNet-Pretrain DDPM, it also can produce reconstructions with a certain degree of fidelity, in stark contrast to outputs from stochastic DDPM sampling, which retain only marginal semantic relevance. In addition, to provide deeper insights into the latent representations, Fig. 4 presents a visualization of the k-means clusters discerned by distinct blocks at different timesteps. The qualitative result across different domains and modalities suggests the usage of these representations for domain-generalizable dense prediction tasks.

Comparison and Ablation Study on the IDRiD Dataset Table 1 presents the quantitative results of each lesion segmentation task. We observe that our DiffDGSS achieves the highest ROC scores across all lesions in comparison with state-of-the-art methods, indicating superior discriminative power. Despite this, there remains room for improvement in PR scores, which is especially pertinent given the challenge of class imbalance that hampers the performance of small lesions like microaneurysms. We also conduct ablation studies in Table 1 to better understand the impact of the major contributions of our DiffDGSS. The ablation result demonstrates the superiority of the representation derived from deterministic inversion over that of stochastic reverse diffusion and the effectiveness of our feature interpreter design in contrast to simple MLPs [1].

Table 2. Quantitative comparisons with state-of-the-art DGSS methods over retinal vessel segmentation. Top 1 results are highlighted in bold.

Method	Cross-Domain					Cross-Modality		
	HRF	CHASE	DRIVE	STARE	Average	ROSE	OCTA	Average
20'BigAug [29]	70.06	76.50	76.42	79.61	75.65	-	-	-
21'FedDG [16]	71.85	76.40	76.61	80.92	76.44	8.51	6.13	7.32
22'AADG [18]	72.57	78.34	77.70	**81.79**	77.60	61.57	50.78	56.18
21'SemanticGAN [14]	-	-	-	-	-	53.99	50.65	52.32
DiffDGSS (Ours)	**74.12**	**78.47**	**77.72**	80.46	**77.69**	**66.88**	**61.99**	**64.43**

Cross-Domain and Cross-Modality Generalization on Retinal Vessel Segmentation We adopt the leave-one-domain-out strategy [18] to evaluate the performance of DGSS methods across four distinct domains, with the comparative results detailed in Table 2. We observe that DiffDGSS outperforms the

alternatives, leading in HRF, CHASE, and DRIVE datasets with the highest average DSC score of 77.69. To further corroborate the generalization ability of DiffDGSS, we extend our evaluation to include a cross-modality experiment on two OCTA datasets, achieving an impressive 64.43 average DSC score. Overall, the results show that DiffDGSS is robust across various domains and modalities.

4 Conclusion

In this paper, by delving into the powerful yet under-explored potential of the latent space in pre-trained DDPM, we introduce a novel framework DiffDGSS for generalizable retinal image segmentation. Our findings demonstrated that diffusion models are inherently effective for DGSS through the meticulous generative modeling of unlabelled data, holding a promise to overcome the persistent data heterogeneity and annotation scarcity in intricate medical image segmentation.

Acknowledgments. This work was supported partly by the National Natural Science Foundation of China (Grant Nos. 62171312, U22A2024, and 62271328), Shenzhen Science and Technology Program (Grant No. JCYJ20220818095809021), Shenzhen Medical Research Funds(No.C2301005), National Natural Science Foundation of Guangdong Province (Nos. 2024A1515011950).

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

References

1. Baranchuk, D., Voynov, A., Rubachev, I., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. In: International Conference on Learning Representations (2022)
2. Budai, A., Bock, R., Maier, A., Hornegger, J., Michelson, G., et al.: Robust vessel segmentation in fundus images. International journal of biomedical imaging **2013** (2013)
3. Carrión, H., Norouzi, N.: Fedd-fair, efficient, and diverse diffusion-based lesion segmentation and malignancy classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 270–279. Springer (2023)
4. Cuadros, J., Bresnick, G.: Eyepacs: an adaptable telemedicine system for diabetic retinopathy screening. Journal of diabetes science and technology **3**(3), 509–516 (2009)
5. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
6. Fraz, M.M., Remagnino, P., Hoppe, A., Uyyanonvara, B., Rudnicka, A.R., Owen, C.G., Barman, S.A.: An ensemble classification-based approach applied to retinal blood vessel segmentation. IEEE Transactions on Biomedical Engineering **59**(9), 2538–2548 (2012)
7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)

8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
9. 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)
10. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
11. Huang, S., Li, J., Xiao, Y., Shen, N., Xu, T.: Rtnet: relation transformer network for diabetic retinopathy multi-lesion segmentation. IEEE Transactions on Medical Imaging **41**(6), 1596–1607 (2022)
12. Hung, W.C., Tsai, Y.H., Liou, Y.T., Lin, Y.Y., Yang, M.H.: Adversarial learning for semi-supervised semantic segmentation. In: 29th British Machine Vision Conference, BMVC 2018 (2019)
13. Kwon, M., Jeong, J., Uh, Y.: Diffusion models already have a semantic latent space. In: International Conference on Learning Representations (2022)
14. Li, D., Yang, J., Kreis, K., Torralba, A., Fidler, S.: Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8300–8311 (2021)
15. Li, M., Zhang, Y., Ji, Z., Xie, K., Yuan, S., Liu, Q., Chen, Q.: Ipn-v2 and octa-500: Methodology and dataset for retinal image segmentation. arXiv preprint arXiv:2012.07261 **5**, 16 (2020)
16. Liu, Q., Chen, C., Qin, J., Dou, Q., Heng, P.A.: Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1013–1023 (2021)
17. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Liu, Y.: Vmamba: Visual state space model. arXiv preprint arXiv:2401.10166 (2024)
18. Lyu, J., Zhang, Y., Huang, Y., Lin, L., Cheng, P., Tang, X.: Aadg: automatic augmentation for domain generalization on retinal image segmentation. IEEE Transactions on Medical Imaging **41**(12), 3699–3711 (2022)
19. Ma, Y., Hao, H., Xie, J., Fu, H., Zhang, J., Yang, J., Wang, Z., Liu, J., Zheng, Y., Zhao, Y.: Rose: a retinal oct-angiography vessel segmentation dataset and new model. IEEE transactions on medical imaging **40**(3), 928–939 (2020)
20. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6038–6047 (2023)
21. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171. PMLR (2021)
22. Playout, C., Duval, R., Cheriet, F.: A novel weakly supervised multitask architecture for retinal lesions segmentation on fundus images. IEEE transactions on medical imaging **38**(10), 2434–2444 (2019)
23. Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Son, J., Bae, W., Liu, L., Wang, J., Liu, X., Gao, L., et al.: Idrid: Diabetic retinopathy–segmentation and grading challenge. Medical image analysis **59**, 101561 (2020)
24. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2020)

25. Staal, J., Abràmoff, M.D., Niemeijer, M., Viergever, M.A., Van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. IEEE transactions on medical imaging **23**(4), 501–509 (2004)
26. Wu, Y., He, K.: Group normalization. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
27. Xia, W., Zhang, Y., Yang, Y., Xue, J.H., Zhou, B., Yang, M.H.: Gan inversion: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 3121–3138 (2022)
28. Yap, B.P., Ng, B.K.: Cut-paste consistency learning for semi-supervised lesion segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6160–6169 (2023)
29. Zhang, L., Wang, X., Yang, D., Sanford, T., Harmon, S., Turkbey, B., Wood, B.J., Roth, H., Myronenko, A., Xu, D., et al.: Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE transactions on medical imaging **39**(7), 2531–2540 (2020)
30. Zhang, Y., Ling, H., Gao, J., Yin, K., Lafleche, J.F., Barriuso, A., Torralba, A., Fidler, S.: Datasetgan: Efficient labeled data factory with minimal human effort. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10145–10155 (2021)
31. Zhou, Y., He, X., Huang, L., Liu, L., Zhu, F., Cui, S., Shao, L.: Collaborative learning of semi-supervised segmentation and classification for medical images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2079–2088 (2019)
32. Zoph, B., Ghiasi, G., Lin, T.Y., Cui, Y., Liu, H., Cubuk, E.D., Le, Q.: Rethinking pre-training and self-training. Advances in neural information processing systems **33**, 3833–3845 (2020)