FairDiff: Fair Segmentation with Point-Image Diffusion

Wenyi Li[*1], Haoran Xu[*1], Guiyu Zhang[*1], Huan-ang Gao[1], Mingju Gao[1], Mengyu Wang[2], and Hao Zhao[† 1]

[1] Institute for AI Industry Research (AIR), Tsinghua University
[2] Harvard Ophthalmology AI Lab, Harvard University
liwenyi19@mails.ucas.ac.cn, zhaohao@air.tsinghua.edu.cn

Abstract. Fairness is an important topic for medical image analysis, driven by the challenge of unbalanced training data among diverse target groups and the societal demand for equitable medical quality. In response to this issue, our research adopts a data-driven strategy—enhancing data balance by integrating synthetic images. However, in terms of generating synthetic images, previous works either lack paired labels or fail to precisely control the boundaries of synthetic images to be aligned with those labels. To address this, we formulate the problem in a joint optimization manner, in which three networks are optimized towards the goal of empirical risk minimization and fairness maximization. On the implementation side, our solution features an innovative Point-Image Diffusion architecture, which leverages 3D point clouds for improved control over mask boundaries through a point-mask-image synthesis pipeline. This method outperforms significantly existing techniques in synthesizing scanning laser ophthalmoscopy (SLO) fundus images. By combining synthetic data with real data during the training phase using a proposed Equal Scale approach, our model achieves superior fairness segmentation performance compared to the state-of-the-art fairness learning models. Code is available at https://github.com/wenyi-li/FairDiff.

Keywords: Fairness Learning · Image Synthesis · Fundus Image Segmentation · Diffusion Models

1 Introduction

The pursuit of fairness in medical image analysis is an important topic because training data for different groups are usually unbalanced while the society calls for equitable medical quality across diverse target groups. To address the issue of unfairness, there are two principal strategies, optimization-driven and data-driven approaches. The optimization-driven approaches [32,16,21,10,18] apply fair learning methods during training over perception models [7,27,8], such as adjusting the weight of loss for different sensitive attributes or incorporating additional regularization losses to minimize bias across groups. By contrast, a

---

⋆ [*] Indicates Equal Contribution. [†] Indicates Corresponding Author.
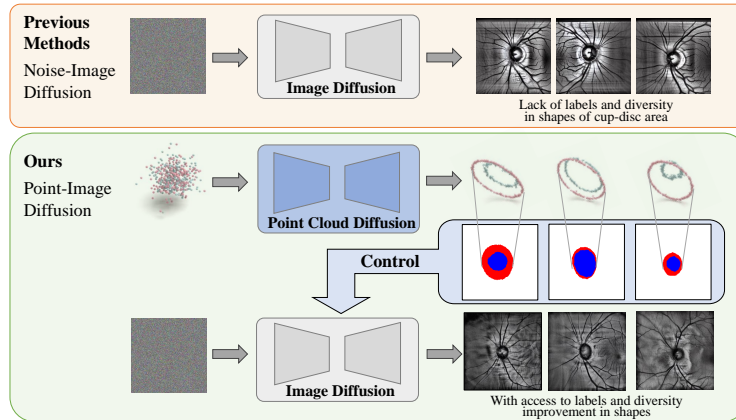
Fig. 1. Comparison of Traditional Noise-Image Diffusion and Our Point-Image Diffusion Methods. We transform 2D mask data into a 3D point cloud format, leveraging the spatial coordinates to delineate boundaries more accurately.

more principled solution, what we name as data-driven approach, addresses the root cause of the issue—unbalanced data distribution through the augmentation of datasets with additional images from underrepresented groups. However, acquiring medical images, such as scanning laser ophthalmoscopy (SLO) fundus images, from vulnerable and under-represented populations proves challenging [22,25]. Therefore, we resort to image synthesis methods [20,33,6] to generate additional data, aiming to improve the fairness of results.

A recent study FairSeg [28] proposes the first fairness dataset for medical segmentation, including 10,000 SLO fundus images from 10,000 patients with pixel-wise disc and cup segmentation mask annotation. The cup-disc area is important for diagnosing a range of eye conditions. However, the anatomical structures of the fundus vary across different racial groups. For instance, Blacks often have a larger cup-to-disc ratio than other races, and Asians are more prone to angle-closure glaucoma than Whites. Therefore, the diversity of the cup-disc area present challenges to the synthesis of images.

Previous medical image synthesis works [14,5,3] mainly focused on the direct generation of medical images. Although these synthesized images often closely mimic the distribution of real images, they lack paired labels, and the process of annotating these images is time-consuming and labor-intensive. Several mask-to-image works, like Freemask [30] and OASIS [26], can generate images from masks but utilize the same set of masks for both real and synthetic samples, which lead to a lack of diversity. SEGGEN [31] proposed MaskSyn which can generate masks through 2D diffusion model [19]. However, accurately controlling these mask boundaries in a two-dimensional space using GANs [9,1,13] and 2D diffusion models [12,20,4] is challenging, because of the inherent constraints in capturing boundaries with pixel-level detail [23,24].
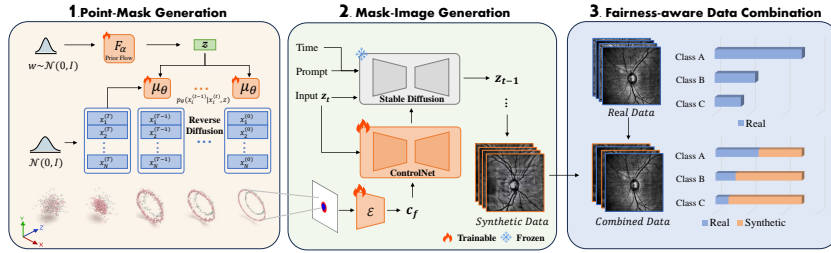
Fig. 2. Overview of Our Fairness-aware Point-Image Diffusion Framework.

To address these challenges, we explore the utilization of sampling points that convert pixel boundaries into spatial coordinates to enhance boundary control. Sampling points on a 2D mask results in a 2D point cloud. However, to distinguish different category boundaries, we transform the point cloud into a 3D format for better point feature learning, where the z-axis is used to mark categories, while the x and y coordinates retain their roles as the plane coordinates. The process is illustrated in Fig. 1.

In this paper, we formulate the fairness problem in a joint optimization manner, in which three networks are optimized towards the goal of empirical risk minimization and fairness maximization. On the implementation side, we introduce a novel Point-Image Diffusion architecture. In this framework, we first generate segmentation masks using point cloud diffusion and then synthesize images based on the control of these synthesized masks. After acquiring a substantial amount of synthetic data for various minority groups, we employ an equal-scale data-combining approach to ensure that the populations of each sensitive attribute group are balanced.

Contributions. In summary, our contributions are as follows: **(1)** We introduce a novel Point-Image diffusion approach for medical image synthesis, which utilizes 3D point clouds to enhance mask boundary control. This method outperforms significantly existing techniques in SLO fundus image synthesis. **(2)** Downstream segmentation tasks verified the efficacy of the synthesized data, demonstrating improvement in model segmentation performance. **(3)** By simply integrating synthetic and real data in the training phase, we achieve superior fairness segmentation performance compared to the state-of-the-art fairness learning models.

2   Methodology

2.1   Overview

Preliminaries. Fairness in a segmentation model can be defined as the model's ability to score evenly for images of different target groups. Consider a dataset D that is made up of image pairs (x, y). Here, x represents the input image and y is the ground truth segmentation mask. $\hat{y}$ refers to the predicted mask. To

measure the fairness of segmentation, we introduce a metric called $\text{Fairness}_\theta(D)$ on the dataset D of segmentation method θ. If we have sensitive attributes $S = \{s_0, ..., s_i, ...s_k\}$, each attribute can divide the population into groups $G = \{g_1, ..., g_n\}$. The fairness metric for segmentation can be represented as:

$$\text{Fairness}_\theta(D) = -\sum_{i=1}^{k} \left(\text{Var}_G \left[\text{M}_\theta(\hat{y}, y|s_i)\right]\right) \tag{1}$$

where $\text{M}_\theta(\hat{y}, y|s_i)$ is the performance metrics such as mIoU or Dice coefficient, Var_G is the variance of metrics within the groups under s_i.

Overview of Framework. To achieve fairness, our research primarily explores data-driven methodologies, focusing on the use of synthetic data. As illustrated in Fig 2, we introduce a novel Point-Image architecture designed specifically for the high-quality SLO fundus images synthesis. The first step is to generate segmentation masks from a Gaussian noise $\mathcal{N}$ using a 3D diffusion model with parameters ϕ, and then to generate images x from the masks using a 2D diffusion model with parameters ψ. After integrating data from real datasets with synthesized data, we proceed to train them using a segmentation model that is parameterized by θ. The overall optimization can be defined as:

$$\min_{\theta} \max_{\phi, \psi} \text{Fairness}(D, \theta, \phi, \psi) \tag{2}$$

2.2   Point-Mask Generation

To produce diverse cup-disc shape fundus images for SLO and acquire precise label maps for synthetic image data, we first augment segmentation masks using the labels from actual real-world data.

Transformation to 3D Point Cloud. Given a 2D mask image of size $W \times H$, where W and H are the width and height of the image. The function $f : I \to P$ maps I to a 3D point cloud P for training. f is defined as follows:

$$f(I) = \{p_i = (x_i, y_i, z_i) \mid (x_i, y_i) \in \text{Boundary}(I),\ z_i = g((x_i, y_i), I)\} \tag{3}$$

where (x_i, y_i) are the coordinates of a pixel in I, $\text{Boundary}(I)$ represents those pixels that are situated at the segmentation boundaries, $g((x_i, y_i), I)$ is a function that assigns the z_i value based on the pixel's position. g is defined as follows:

$$g((x_i, y_i), I) = \begin{cases} z_0 & \text{if } (x_i, y_i) \text{ is on the boundary of the Cup} \\ -z_0 & \text{if } (x_i, y_i) \text{ is on the boundary of the Disc} \end{cases} \tag{4}$$

Point Cloud Diffusion for Generation. After converting the existing 2D labels into 3D point clouds, we employ a point cloud diffusion model based on [15] to learn the distribution of these point clouds. The primary training goal of this model is to simulate the reverse of a random diffusion process, learning to move from a normal distribution to the distribution of real point clouds. During

the training phase, we introduce varying degrees of random noise into the point clouds and ensure that the denoising model predicts noise that closely matches the actual noise added.

For each group g_i of the sensitive attribute S, we train a point cloud diffusion model ϕ_i. Since ϕ_i can effectively capture the characteristics of the cup-disc contour for different groups, we can selectively augment samples for different groups, particularly for minority populations. Through this approach, we prepare the label sets for subsequent procedure.

2.3 Mask-Image Generation

Given the generated mask m as condition c, the next step is to synthesize an image x . This integration of m into the neural network is achieved by introducing an extra condition c into a neural network block, via an architecture known as ControlNet [33]. This method involves freezing the original Stable Diffusion block's parameters Φ, replicating it into a trainable copy with parameters Φ_c, and connecting these blocks with two zero-initialized 1×1 convolutional layers. Specifically, the mask m is encoded into tokens $c_f = E(c_i)$, which are then fed into ControlNet. The output of ControlNet y_c is given by:

$$y_c = F(x; \Phi) + Z\left(F\left(x + Z(c; \Phi_{z1}); \Phi_c\right); \Phi_{z2}\right) \tag{5}$$

where y_c is the output from the ControlNet block, $Z(\cdot; \cdot)$ represents the zero convolutional layers, and Φ_{z1} and Φ_{z2} are the parameters of the two zero convolutional layers. At the beginning of training, y_c equals y due to the zero initialization of the zero convolutional layers' weights and biases, ensuring no harmful noise is introduced into the network's hidden states. As training progresses, the zero convolutional layers gradually adapt the output based on the input condition c_f, thereby achieving control over the original feature map x.

2.4 Equal-Scale Data Combination

The method of combining real and synthetic data is straightforward, termed Equal-Scale Data Combination, which balances the sample sizes across all sensitive groups. Assume $D_r = \{x_{r,1}, x_{r,2}, \ldots, x_{r,N_r}\}$ and $D_s = \{x_{s,1}, x_{s,2}, \ldots, x_{s,N_s}\}$ as the sets of sample points from the real data distribution P_r and the synthetic data distribution P_s, respectively. Here, N_r and N_s denote the number of samples in the real and synthetic datasets, respectively. The equal-scale combination process involves augmenting the dataset with synthetic samples for underrepresented groups or possibly subsampling overrepresented groups. For each sensitive group g, if $N_{g,r} < N_{target}$, generate $(N_{target} - N_{g,r})$ synthetic samples from P_s specific to group g, resulting in a combined set $D_{g,s}^*$. If $N_{g,r} > N_{target}$, random sample N_{target} samples from $D_{g,r}^*$. N_{target} is the target sample size, which could be based on the size of the largest group, a specified threshold for fairness. The combined dataset D for training can be represented as:

$$D = \bigcup_g (D_{g,r}^* \cup D_{g,s}^*) \tag{6}$$

3   Experiments and Results

3.1   Setup

Datasets. We use Harvard-FairSeg [28] as the real SLO fundus image dataset. It includes six critical attributes for comprehensive studies on fairness, which are age, gender, race, ethnicity, language preference, and marital status. The fairness and segmentation results of all models, whether using synthetic data or not, are tested on the test split of 2,000 real SLO fundus images.

Segmentation Models. To verify the impact of our synthetic data on segmentation and fairness, we selected two segmentation models, including a small model TransUNet [2] and a larger model SAMed [34] (the experiments on the latter are provided in the supplementary).

Training Details. For the training of image synthesis, we utilize 512 points to sample the boundaries of each original mask. These point clouds are then normalized. Training is performed on NVIDIA 3090 GPUs with a batch size of 48 and a learning rate of 1e-4 across 2,000,000 steps. For the training of the segmentation model, following the experimental setup of Fairseg [28], we employed a combination of cross-entropy and Dice losses as the training loss and used the AdamW optimizer. To enable effective comparisons, TransUNet was trained with a base learning rate of 0.01 and a momentum of 0.9 over 300 epochs, while SAMed was set with a base learning rate of 0.005 and a momentum of 0.9, undergoing training for 100 epochs. The batch size for both was set to 48. For the number of training samples, we have controlled it to be 8000, whether using all real data or a mix of real and synthetic data.

3.2   Synthesis Quality results

Metrics. To evaluate the generation quality, we employ the Fréchet inception distance (FID) [11], minimum matching distance (MMD) and the coverage score (COV). The detailed definitions of these metrics can be found in the *supplementary materials.*

Results. We compare our Point-Image image generation method with several state-of-the-art methods, including Stable Diffusion 1.5 [20], pix2pixHD [29], OASIS [26], SPADE [17] and ControlNet [33]. As shown in Tab. 1, our method significantly outperforms existing techniques in SLO fundus image synthesis. Notably, our approach achieves the lowest FID score, indicating that our generated images bear a closer resemblance to the actual images when compared to other methods. Furthermore, the MMD results suggest that our method also more accurately replicates the distribution of the original image dataset.

Ablation over two-stage diffusion. Comparing with ControlNet [33] (one-stage label-to-image synthesis), our two-stage pipeline, where we first sample labels and then synthesis images, shows effectiveness in generating diverse images, as reflected by the highest COV (Coverage) score among the methods evaluated. The enhancement in image quality and diversity underscores the efficacy of our image synthesis technique. Fig. 3 visualizes the results of image synthesis.
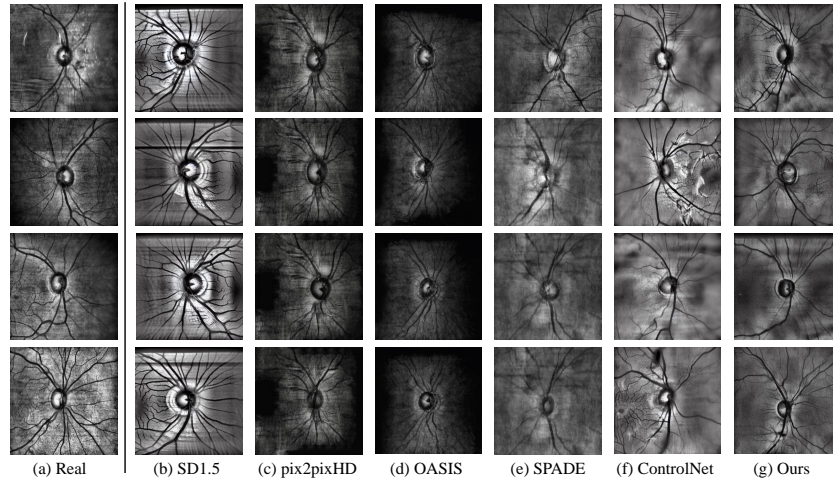
(a) Real   (b) SD1.5   (c) pix2pixHD   (d) OASIS   (e) SPADE   (f) ControlNet   (g) Ours

Fig. 3. Visualization Results of Different Image Synthesis Results.

Table 1. Comparison of Synthesis Quality.

Method	FID↓	MMD↓	COV↑
SD1.5[20]	167.39	33.21	3.13
pix2pixHD[29]	157.02	22.73	4.63
OASIS[26]	89.92	28.57	3.07
SPADE[17]	77.26	23.82	5.75
ControlNet[33] (w/o Point-Mask)	67.29	23.60	9.45
Ours (w/ Point-Mask)	**60.51**	**20.06**	**10.83**

3.3 Fairness Segmentation Results

Metrics. Following prior work [28], we measure the fairness segmentation results using Equity-Scaled Segmentation Performance (ESSP), which is defined as

$$\text{ESSP} = \frac{\mathcal{L}(\{(z', y)\})}{1 + \text{Stdev}} \tag{7}$$

where $\mathcal{L}$ is the Dice or IoU metric. The ES-Dice and ES-IoU metrics consider both segmentation performance and fairness across all groups. The conventional Overall Dice and IoU metrics only assess the segmentation performance.

Results. In our comparative analysis, we examine the performance of our Equal Scale Data Combination method against several state-of-the-art fairness-learning approaches, including ADV [16], GroupDRO [21], and FairSeg [28]. This evaluation encompasses experiments conducted across four sensitive attributes. The detailed results for TransUNet are presented from Table 2 to Table 5. Due to limitations in space, the results of SAMed are included in the supplementary materials. From the perspective of racial fairness, Tab. 2 highlights our Equal

Table 2. TransUNet segmentation performance of Optic Cup and Rim (**Sensitive attribute = Race**)

	Method	Overall ES-Dice↑	Overall Dice↑	Overall ES-IoU↑	Overall IoU↑	Asian Dice↑	Black Dice↑	White Dice↑	Asian IoU↑	Black IoU↑	White IoU↑
Cup	TransUNet	0.8372	**0.8481**	0.7409	**0.7532**	0.8270	**0.8489**	0.8503	0.7277	**0.7576**	0.7551
	TransUNet+ADV	0.8325	0.8410	0.7345	0.7432	0.8246	0.8417	0.8426	0.7260	0.7482	0.7440
	TransUNet+GroupDRO	0.8313	0.8442	0.7359	0.7479	0.8197	0.8469	0.8464	0.7232	0.7529	0.7495
	TransUNet+FairSeg	0.8350	0.8464	0.7374	0.7497	0.8248	0.8484	0.8484	0.7247	0.7550	0.7513
	Ours(Equal Scale)	**0.8397**	0.8480	**0.7441**	0.7529	**0.8320**	0.8483	0.8497	**0.7352**	0.7572	0.7540
Rim	TransUNet	0.7604	0.7927	0.6393	0.6706	0.7457	0.7307	0.8106	0.6160	0.5991	0.6913
	TransUNet+ADV	0.7579	0.7906	0.6371	0.6682	0.7413	0.7286	0.8087	0.6116	0.5982	0.6888
	TransUNet+GroupDRO	0.7564	0.7896	0.6351	0.6674	0.7470	0.7229	0.8080	0.6183	0.5899	0.6887
	TransUNet+FairSeg	0.7628	0.7950	0.6410	0.6725	0.7479	0.7325	0.8130	0.6185	0.6020	0.6935
	Ours(Equal Scale)	**0.7697**	**0.7999**	**0.6494**	**0.6797**	**0.7565**	**0.7427**	**0.8165**	**0.6279**	**0.6114**	**0.6994**

Table 3. TransUNet segmentation performance of Optic Cup and Rim (**Sensitive attribute = Gender**)

	Method	Overall ES-Dice↑	Overall Dice↑	Overall ES-IoU↑	Overall IoU↑	Male Dice↑	Female Dice↑	Male IoU↑	Female IoU↑
Cup	TransUNet	0.8448	0.8481	0.7502	0.7532	0.8458	0.8513	0.7508	0.7564
	TransUNet+ADV	0.8343	0.8345	0.7351	0.7356	0.8344	0.8348	0.7361	0.7350
	TransUNet+GroupDRO	0.8426	0.8478	0.7473	0.7522	0.8441	0.8528	0.7483	0.7575
	TransUNet+FairSeg	**0.8477**	0.8489	0.7502	0.7530	**0.8494**	0.8514	0.7505	0.7556
	Ours(Equal Scale)	0.8461	**0.8505**	**0.7522**	**0.7564**	0.8474	**0.8548**	**0.7531**	**0.7610**
Rim	TransUNet	0.7895	0.7927	0.6673	0.6706	0.7951	0.7894	0.6736	0.6665
	TransUNet+ADV	0.7783	0.7852	0.6553	0.6630	0.7905	0.7779	0.6699	0.6534
	TransUNet+GroupDRO	0.7901	0.7917	0.6681	0.6699	0.7930	0.7900	0.6716	0.6677
	TransUNet+FairSeg	0.7893	0.7898	0.6698	0.6698	0.7924	0.7932	0.6678	0.6653
	Ours(Equal Scale)	**0.7945**	**0.7981**	**0.6745**	**0.6780**	**0.8007**	**0.7944**	**0.6811**	**0.6737**

Table 4. TransUNet segmentation performance of Optic Cup and Rim (**Sensitive attribute = Language**)

	Method	Overall ES-Dice↑	Overall Dice↑	Overall ES-IoU↑	Overall IoU↑	English Dice↑	Spanish Dice↑	Others Dice↑	English IoU↑	Spanish IoU↑	Others IoU↑
Cup	TransUNet	0.8255	0.8481	0.7273	0.7532	0.8469	0.8972	0.8531	0.7516	**0.8166**	0.7592
	TransUNet+ADV	0.8071	0.8312	0.7056	0.7323	0.8296	0.8833	0.8338	0.7301	0.7990	0.7376
	TransUNet+GroupDRO	0.8231	0.8416	0.7231	0.7442	0.8398	0.8844	0.8571	0.7421	0.7993	0.7605
	TransUNet+FairSeg	0.8277	0.8481	0.7289	0.7523	0.8467	**0.8934**	0.8562	0.7504	0.8109	0.7619
	Ours(Equal Scale)	**0.8358**	**0.8497**	**0.7353**	**0.7542**	**0.8479**	0.8809	**0.8686**	**0.7519**	0.8033	**0.7755**
Rim	TransUNet	0.7721	0.7927	0.6525	0.6706	0.7940	**0.8165**	0.7633	0.6721	0.6950	0.6398
	TransUNet+ADV	0.7690	0.7884	0.6501	0.6666	0.7903	0.7964	0.7501	0.6687	0.6717	0.6265
	TransUNet+GroupDRO	0.7691	0.7857	0.6468	0.6613	0.7867	0.8057	0.7628	0.6625	0.6800	0.6355
	TransUNet+FairSeg	0.7725	0.7898	0.6524	0.6668	0.7909	0.8106	0.7661	0.6680	0.6865	0.6424
	Ours(Equal Scale)	**0.7783**	**0.7959**	**0.6578**	**0.6743**	**0.7970**	0.8161	**0.7711**	**0.6755**	**0.6969**	**0.6469**

Scale method's effectiveness, achieving the highest ES-Dice in both the Cup and Rim area among all racial groups Asian, Black, and White, with remarkable scores of 0.8397 and 0.7697, respectively. The ES-IoU metric also supports our method, highlighting its effectiveness in achieving both accurate and equitable segmentation. Tab. 3, Tab. 4 and Tab. 5 also demonstrate the capability to enhance fairness metrics (ES-Dice & ES-IoU) and segmentation performance (Dice & IoU).

Table 5. TransUNet segmentation performance of Optic Cup and Rim **(Sensitive attribute = Ethnicity)**

	Method	Overall ES-Dice↑	Overall Dice↑	Overall ES-IoU↑	Overall IoU↑	Hispanic Dice↑	Non-Hispanic Dice↑	Hispanic IoU↑	Non-Hispanic IoU↑
Cup	TransUNet	0.8339	0.8481	0.7366	0.7532	0.8463	**0.8704**	0.7508	**0.7826**
	TransUNet+ADV	0.8171	0.8320	0.7149	0.7315	0.8304	0.8561	0.7294	0.7622
	TransUNet+GroupDRO	0.8376	0.8482	0.7406	0.7526	0.8468	0.8648	0.7507	0.7735
	TransUNet+FairSeg	0.8388	**0.8483**	**0.7412**	**0.7542**	**0.8501**	0.8661	0.7515	0.7764
	Ours(Equal Scale)	**0.8439**	0.8462	0.7402	0.7500	0.8485	0.8448	**0.7664**	0.7477
Rim	TransUNet	0.7848	0.7927	0.6650	0.6706	0.7914	**0.8057**	0.6695	**0.6815**
	TransUNet+ADV	0.7793	0.7841	0.6570	0.6602	0.7829	0.7915	0.6590	0.6658
	TransUNet+GroupDRO	0.7924	0.7943	0.6694	**0.6733**	0.7936	0.7901	0.6728	0.6646
	TransUNet+FairSeg	0.7884	**0.7939**	**0.6710**	0.6754	**0.7943**	0.8040	**0.6697**	0.6789
	Ours(Equal Scale)	**0.7886**	0.7902	0.6666	0.6670	0.7917	0.7888	0.6649	0.6657

4   Conclusion

In this study, we analyze fairness within the context of medical image segmentation and address the challenge of data imbalance through the use of synthetic data. We present a novel Point-Image Diffusion method tailored for synthesizing SLO fundus images, which significantly outperforms existing techniques in this domain. By incorporating both synthetic and real data during the training phase utilizing the Equal Scale method, we achieve a comprehensive improvement in both accuracy and fairness across various sensitive attributes.

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

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning. pp. 214–223. PMLR (2017)
2. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
3. Dalmaz, O., Yurt, M., Çukur, T.: Resvit: Residual vision transformers for multi-modal medical image synthesis. IEEE Transactions on Medical Imaging **41**(10), 2598–2614 (2022)
4. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
5. Friedrich, P., Wolleb, J., Bieder, F., Durrer, A., Cattin, P.C.: Wdm: 3d wavelet diffusion models for high-resolution medical image synthesis. arXiv preprint arXiv:2402.19043 (2024)
6. Gao, H.a., Gao, M., Li, J., Li, W., Zhi, R., Tang, H., Zhao, H.: Scp-diff: Photo-realistic semantic image synthesis with spatial-categorical joint prior. arXiv preprint arXiv:2403.09638 (2024)
7. Gao, H.a., Tian, B., Li, P., Chen, X., Zhao, H., Zhou, G., Chen, Y., Zha, H.: From semi-supervised to omni-supervised room layout estimation using point clouds. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 2803–2810. IEEE (2023)
8. Gao, H.a., Tian, B., Li, P., Zhao, H., Zhou, G.: Dqs3d: Densely-matched quantization-aware semi-supervised 3d detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21905–21915 (2023)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
10. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)
11. 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)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
13. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
14. Khader, F., Mueller-Franzes, G., Arasteh, S.T., Han, T., Haarburger, C., Schulze-Hagen, M., Schad, P., Engelhardt, S., Baessler, B., Foersch, S., et al.: Medical diffusion–denoising diffusion probabilistic models for 3d medical image generation. arXiv preprint arXiv:2211.03364 (2022)
15. Luo, S., Hu, W.: Diffusion probabilistic models for 3d point cloud generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2837–2845 (2021)
16. Madras, D., Creager, E., Pitassi, T., Zemel, R.: Learning adversarially fair and transferable representations. In: International Conference on Machine Learning. pp. 3384–3393. PMLR (2018)
17. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2337–2346 (2019)

18. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., Weinberger, K.Q.: On fairness and calibration. Advances in neural information processing systems **30** (2017)
19. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
20. 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. pp. 10684–10695 (2022)
21. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. arXiv preprint arXiv:1911.08731 (2019)
22. Sharma, A., Palaniappan, L.: Improving diversity in medical research. Nature Reviews Disease Primers **7**(1), 74 (2021)
23. Shen, J., Lu, S., Qu, R., Zhao, H., Zhang, L., Chang, A., Zhang, Y., Fu, W., Zhang, Z.: A boundary-guided transformer for measuring distance from rectal tumor to anal verge on magnetic resonance images. Patterns **4**(4) (2023)
24. Shen, J., Lu, S., Qu, R., Zhao, H., Zhang, Y., Chang, A., Zhang, L., Fu, W., Zhang, Z.: Measuring distance from lowest boundary of rectal tumor to anal verge on ct images using pyramid attention pooling transformer. Computers in Biology and Medicine **155**, 106675 (2023)
25. Shepherd, V.: An under-represented and underserved population in trials: methodological, structural, and systemic barriers to the inclusion of adults lacking capacity to consent. Trials **21**(1), 1–8 (2020)
26. Sushko, V., Schönfeld, E., Zhang, D., Gall, J., Schiele, B., Khoreva, A.: You only need adversarial supervision for semantic image synthesis. arXiv preprint arXiv:2012.04781 (2020)
27. Tian, B., Liu, M., Gao, H.a., Li, P., Zhao, H., Zhou, G.: Unsupervised road anomaly detection with language anchors. In: 2023 IEEE international conference on robotics and automation (ICRA). pp. 7778–7785. IEEE (2023)
28. Tian, Y., Shi, M., Luo, Y., Kouhana, A., Elze, T., Wang, M.: Fairseg: A large-scale medical image segmentation dataset for fairness learning using segment anything model with fair error-bound scaling. In: International Conference on Learning Representations (ICLR) (2024)
29. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018)
30. Yang, L., Xu, X., Kang, B., Shi, Y., Zhao, H.: Freemask: Synthetic images with dense annotations make stronger segmentation models. Advances in Neural Information Processing Systems **36** (2024)
31. Ye, H., Kuen, J., Liu, Q., Lin, Z., Price, B., Xu, D.: Seggen: Supercharging segmentation models with text2mask and mask2img synthesis. arXiv preprint arXiv:2311.03355 (2023)
32. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: International conference on machine learning. pp. 325–333. PMLR (2013)
33. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
34. Zhang, Y., Shen, Z., Jiao, R.: Segment anything model for medical image segmentation: Current applications and future directions. Computers in Biology and Medicine p. 108238 (2024)