Derm-FairAnon: Mitigating Demographic Bias in Skin Image Anonymization with Self-supervised Preference Optimization

Yeon Gyu Han[1,2] Jung Im Na[3] Seong Hwan Kim[4] Dongheon Lee[1,5,6 *]
[1] MODULABS [2] Department of Biomedical Engineering, Chungnam National University
[3] Department of Dermatology, Seoul National University Bundang Hospital
[4] Department of Plastic and Reconstructive Surgery, Kangnam Sacred Hospital
[5] Department of Radiology, Seoul National University College of Medicine
[6] Department of Radiology, Seoul National University Hospital
dusrb37@gmail.com vividna@gmail.com kalosmanus@naver.com
dhlee.jubilee@gmail.com

Abstract

Medical image anonymization faces the critical challenge of balancing patient privacy protection, clinical feature preservation, and demographic fairness. Existing methods often compromise privacy, obscure essential disease information, or perpetuate demographic biases in the anonymized outputs. We propose " ***Derm-FairAnon****" a comprehensive framework for dermatological image anonymization that addresses these challenges through a novel integration of Stable Diffusion-v2 Inpainting with two key contributions: (1) Self-Supervised Preference Optimization (SelfPO), a novel approach that eliminates the need for explicit preference labels by leveraging image augmentation to generate self-supervised ranking signals; and (2) a demographic fairness mechanism with Skin-Fair loss, $\mathcal{L}_{Skin-Fair}$ that enables balanced demographic representation in generated images, effectively mitigating attribute biases while maintaining clinical utility. Evaluated on dermatological images from multiple hospitals, Derm-FairAnon outperforms existing methods in disease classification performance, anonymization success, demographic bias reduction, and clinical assessment by dermatologists.*

1. Introduction

Medical image anonymization faces the critical challenge of balancing patient privacy protection with clinical feature preservation, particularly in dermatology where patient faces and skin lesions often appear in the same frame [1-3]. Traditional anonymization methods applying digital masks or blurring [4, 5] either compromise privacy or obscure disease information on facial regions, resulting in substantial loss of clinically valuable data. GAN-based anonymization methods [6, 7] generate artifacts at boundary regions and fail to accurately represent disease characteristics.

Diffusion models have demonstrated promising capabilities in medical image synthesis [8, 9], but still struggle with key limitations: they cannot reliably preserve disease-specific characteristics (such as scaling patterns in psoriasis versus erythematous patches in atopic dermatitis), and they create inconsistent transitions between preserved pathology and anonymized regions, often inheriting and amplifying demographic biases from training data [8].

The challenge is further complicated by demographic biases in medical image generation. Although face generation and editing can be used for anonymization, most work has relied on datasets with limited demographic diversity, such as CelebA [10] and FFHQ [11]. Consequently, these models fail to accurately represent diverse facial features, particularly those of underrepresented demographic groups. This leads to systematic disparities in anonymization quality across different demographic attributes, resulting in inconsistent performance when applied to diverse clinical populations [8, 12, 13].

We introduce **Derm-FairAnon**, a framework based on Stable Diffusion v2 Inpainting [14] with two key innovations. First, Self-Supervised Preference Optimization (SelfPO) [15] eliminates the need for human feedback by leveraging image augmentation hierarchies ("original > slightly degraded > heavily degraded"), maintaining high-quality disease representations without costly annotation. Second, Skin-Fair loss combines distributional alignment for balanced demographic representation, semantic preservation for maintaining image structure, and diagnostic preservation for protecting critical disease features during bias mitigation.

We evaluated our model using comprehensive clinical skin disease datasets and validated its generalization on

* Corresponding Author.

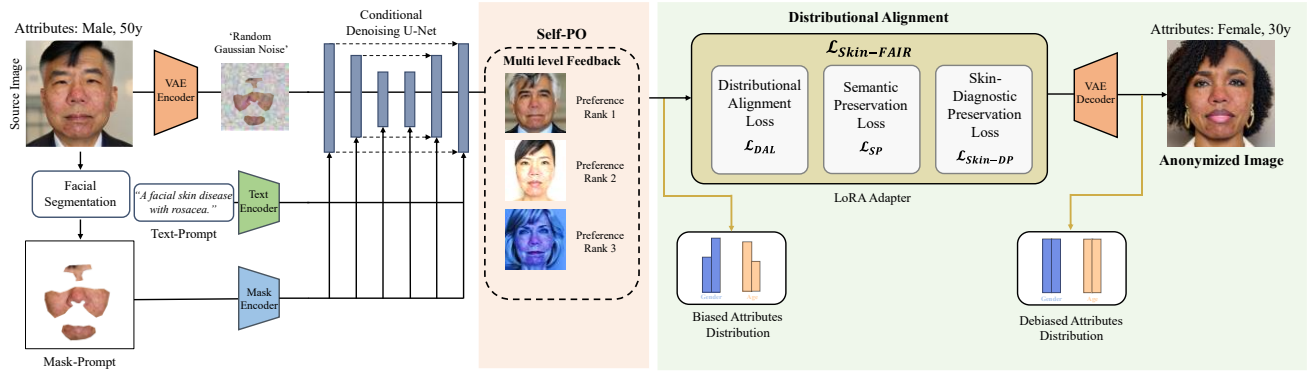

Figure 1: The overview of ***Derm-FairAnon***. The disease-preserving facial segmentation model generates masks that designate which clinical features to preserve during anonymization. Our Self-PO approach creates a multi-level preference structure through controlled image degradation, enabling the model to learn high-quality disease representation without human feedback. The Skin-Fair loss integrates three components: Distributional Alignment Loss to balance demographic attributes, Semantic Preservation Loss to maintain image structure, and Skin-Diagnostic Preservation Loss to protect critical disease features. This comprehensive approach, implemented through LoRA adapters, transforms demographically biased representations into balanced outputs while preserving diagnostic value.

CelebA-HQ and FFHQ test sets. Results demonstrate that Derm-FairAnon significantly outperforms existing methods in clinical utility and feature preservation accuracy, as assessed by board-certified dermatologists, while effectively anonymizing patient identity across diverse skin conditions and demographics.

Our main contributions are: (1) Self-Supervised Preference Optimization that enables disease-aware anonymization without human feedback through strategic self-supervision, and (2) Skin-Fair loss that maintains balanced demographic representations while preserving diagnostic information, avoiding the clinical feature compromise common in previous bias mitigation approaches.

2. Related Work

Medical Face Anonymization. Traditional approaches to medical image anonymization have primarily relied on basic image processing techniques. Pixelation and blurring [16] are widely used in clinical publications but often compromise the clinical value of dermatological images by obscuring disease features alongside identifying characteristics. Studies have shown that these methods still pose privacy risks, as machine learning models can sometimes recover identity information from blurred images [16, 17]. More advanced methods employ digital masking techniques [1] that overlay black rectangles on specific facial features have become standard practice in many medical journals. While these methods preserve some clinical information, they fundamentally cannot maintain disease features in masked regions, creating an inherent trade-off between privacy and clinical utility. Additionally, research has demonstrated that partial masking often fails to prevent re-identification, especially when combined with other available information [17]. Recent advances in deep learning have led to GAN-based anonymization methods [18, 19] that replace real faces with synthetic ones. While promising for general medical imaging, these approaches were not designed for dermatological applications and typically replace the entire facial region, eliminating valuable disease information. They also often produce artifacts at the boundaries between preserved and generated regions [16], particularly problematic in dermatology where boundary characteristics are diagnostically significant.

Fairness in Skin Image Generation. Skin disease image generation models exhibit significant demographic biases. Popular generative models heavily favor light skin tones and male representations [20]. Several approaches have attempted to address these issues: DermDiff [12] and FairSkin [8] mitigated racial bias through specialized text prompts and resampling strategies, while asymmetric quality bias related to skin tone in GANs has been analyzed [18]. Despite these advances, these approaches primarily focused on skin tone while insufficiently addressing other attributes like gender and age. DermDiT [13] employed vision-language models to reduce diagnostic bias but remained primarily focused on skin color variations. Performance gaps between genders have been reduced without addressing age-related biases [21,22]. While FEDD [23] demonstrated consistent performance across diverse skin tones, it lacked explicit fairness mechanisms for gender or age dimensions. The field still lacks integrated approaches that provide fairness across multiple demographic attributes and their intersections in dermatological contexts, particularly methods that balance demographic representation while preserving critical disease characteristics during anonymization.

3. Method

Our framework provides a comprehensive approach to

dermatological image anonymization, designed to simultaneously achieve three competing objectives: privacy protection, clinical utility, and demographic fairness. Our methodology consists of three key components: (1) disease-preserving segmentation to differentiate diagnostically important regions from personally identifiable areas; (2) Self-supervised Preference Optimization (SelfPO) to learn high-quality anonymization that maximally preserves disease characteristics without external feedback; and (3) bias mitigation to generate balanced outputs across diverse demographic attributes. Figure 1 illustrates our overall approach.

3.1. Disease-Preserving Segmentation

For precise disease feature preservation, we employ PointRend [24] trained on 17,697 annotated facial images from A hospital (mIoU: 0.92). This approach efficiently segments both facial features and lesion boundaries—critical for dermatological diagnosis—creating masks that designate which disease regions to preserve while anonymizing identifiable facial features.

3.2. Self-supervised Preference Optimization for Diffusion Model

Existing preference optimization techniques require human preference labels or external reward models. To overcome these limitations, we propose Self-supervised Preference Optimization (SelfPO), a diffusion model-specific approach that automatically generates preference learning signals through systematic image augmentations.

The key insight of SelfPO is to generate natural preference rankings by intentionally injecting hierarchical degradations into image quality. Given an original image I_0, we apply image augmentations of varying intensities to form a quality spectrum. When these transformed images are input to the model, the generated responses exhibit quality differences, forming an intrinsic preference ranking $y_0 > y_1 > y_2$.

We extend traditional Direct Preference Optimization [25] to a multi-level preference structure to better capture the nuanced quality characteristics of dermatological images. Our Multi-level DPO loss function is defined as:

$$\mathcal{L} = -\sum_{i<j} E\left[log\sigma\left(\beta\left(log\pi(y_i) - log\pi(y_j)\right)\right)\right] \quad (1)$$

where x = i represents the input image, π_θ is the model being trained, σ is the sigmoid function, and β is a temperature scaling parameter.

For image transformations, we selected clinically valid transformations including Gaussian noise, Gaussian blur, contrast adjustment, and color jitter. These transformations are applied at two intensity levels (weak and strong) to create quality differences while maintaining diagnostic viability.

3.3. Bias Mitigation

To address demographic biases in dermatological image generation, we develop the Skin-Fair loss ($\mathcal{L}_{Skin-Fair}$) with three key components. First, we introduce a Distributional Alignment loss, $\mathcal{L}_{DA}$ that aligns attribute distributions in generated images with predefined target distributions:

$$\mathcal{L}_{DA} = \sum_{i,j}|freq(i) - target(i)| \quad (2)$$

where index i represent different attribute groups (gender, age, skin tone), and target(i) represents the clinically-appropriate target distribution for each attribute group.

Second, we apply a Semantic Preservation loss, $\mathcal{L}_{SP}$ [26] to ensure that structural integrity of images is not compromised during attribute adjustments. This loss measures cosine distances in CLIP and DINO feature spaces between our generated images and images from the original model generated with identical prompts.

Third, we introduce the Skin Diagnostic Preservation loss $\mathcal{L}_{Skin-DP}$ to preserve diagnostically important information:

$$\mathcal{L} = \lambda_1\mathcal{L}_{feature} + \lambda_2\mathcal{L}_{perceptual} + \lambda_3\mathcal{L}_{diagnostic} \quad (3)$$

The $\mathcal{L}_{diagnostic}$ component uses a DenseNet-121 model specialized for skin disease classification to ensure disease characteristics are preserved during bias mitigation.

To efficiently implement the Skin-Fair mechanism, we adopt a LoRA adapter-based approach, requiring less than 1% additional parameters while maintaining flexibility to independently control various demographic factors.

4. Experiments

4.1. Experimental Settings

Baseline. (1) traditional methods (blurring, masking), (2) face anonymization models (DeepPrivacy [27], AnonFaces [28]), and (3) ablated versions of our model.

Datasets. We evaluated our approach on 6,000 facial skin disease images from three hospitals, encompassing five common conditions: psoriasis, atopic dermatitis, acne, rosacea, and seborrheic dermatitis. The dataset includes diverse demographic attributes: gender, age groups. We used an 8:1:1 train-validation-test split while ensuring demographic balance. To verify generalization capability, we additionally evaluated our models on standard benchmark datasets CelebA-HQ and FFHQ.

Evaluation. Diagnostic preservation was quantified through Disease Classification Performance (DCP) using a DenseNet-121 [29] classifier to compare AUC scores before and after anonymization. Anonymization Success Accuracy was measured using the InsightFace [30] face recognition model. Image quality was measured using FID and PSNR metrics. Demographic fairness was evaluated through three bias metrics: Gender Bias, Age Bias, and Gender-Age Intersectional Bias, measuring frequency disparities between demographic groups. Clinical validation was conducted by two board-certified dermatologists evaluating disease feature preservation and image naturalness on a 5-point scale (1: poor - 5: excellent).

4.2. Comparison with the Baselines

Table 1 presents our approach outperforming baselines in both disease preservation, image quality and Anonymization Success Accuracy. Clinical validation confirmed 4.5 of our anonymized images maintained

Methods	D. AUC ↑	Ano. Acc ↑	FID ↓	PSNR ↑	G. Bias ↓	A. Bias ↓	G x A. Bias ↓	Clinical Score ↑
Blurring	54.8	0.89	157.9	19.3	-	-	-	0.5
Masking	61.2	1.00	183.4	17.8	-	-	-	1.0
DeepPrivacy	72.3	0.98	99.8	25.7	-	-	-	0.5
AnonFaces	68.5	0.97	101.3	26.2	-	-	-	0.5
SD-v2-I	83.2	1.00	134.7	28.3	0.36	0.33	0.40	2.5
DL-I	82.7	1.00	94.4	27.9	0.35	0.34	0.39	2.0
SD-XL-I	85.4	1.00	95.2	27.6	0.38	0.35	0.41	2.5
Ours	**94.7**	**1.00**	**94.9**	**30.3**	**0.15**	**0.14**	**0.13**	**4.5**

Table 1: Comparison with the baseline methods.

D: Disease, Ano: Anonymization, G: Gender, A: Age, SD-v2-I: Stable Diffusion-v2 Inpainting, DL-I: DreamLike Inpainting, SD-XL-I: Stable Diffusion XL Inpainting

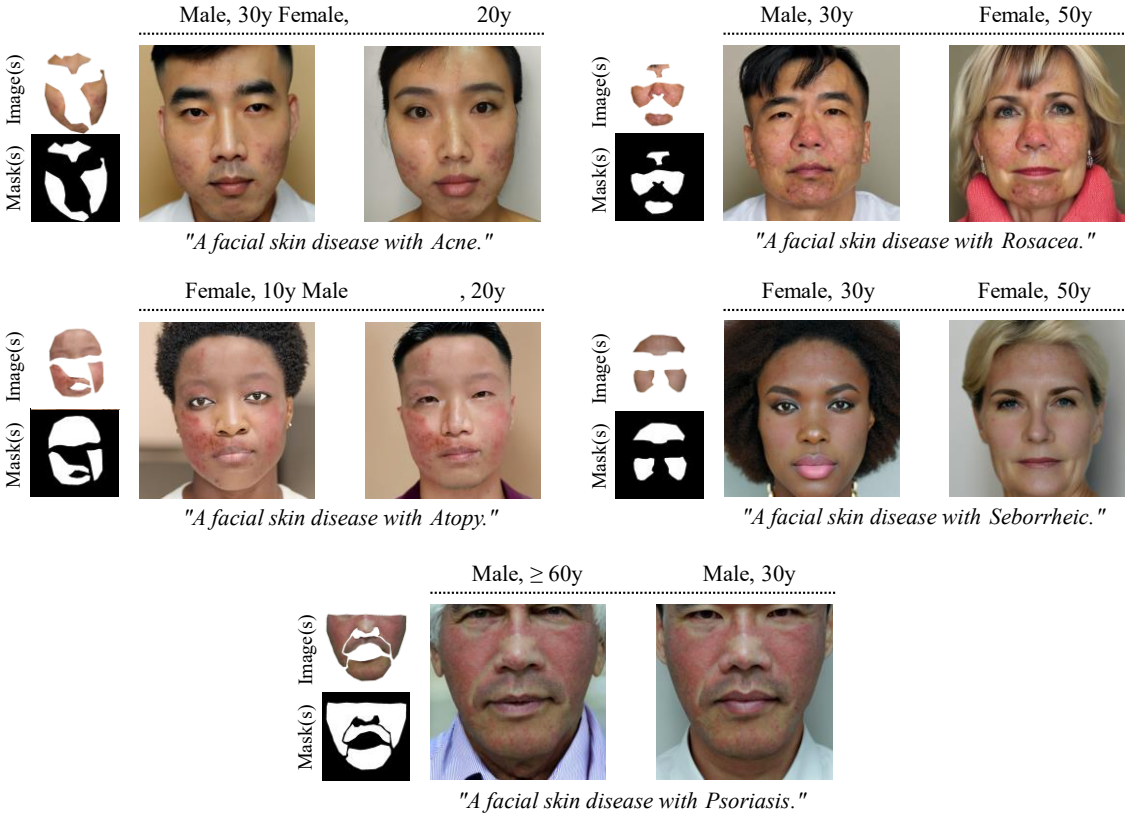

Figure 2: Qualitative results of Derm-FairAnon across diverse skin conditions and demographics. Our approach effectively anonymizes facial images while preserving disease-specific features across different age groups (10y to ≥60y) and genders. The results demonstrate the model's ability to maintain balanced demographic representation while accurately preserving condition-specific characteristics for various dermatological conditions: acne, rosacea, atopic dermatitis, seborrheic dermatitis, and psoriasis.

diagnostic equivalence to originals. Figure 2 shows Quantitative results of Dermatological Image anonymization.

5. Conclusion

We presented ***Derm-FairAnon*** that integrates SelfPO and $\mathcal{L}_{Skin-Fair}$ to achieve high-quality anonymization without human feedback while ensuring balanced demographic representation without compromising diagnostic value. This work represents an important step toward ethical medical data sharing without compromising clinical value.

Acknowledgments

This research was supported by Brian Impact Foundation, a non-profit organization dedicated to the advancement of science and technology for all.

References

[1] Y. Yang, S. Irvine, N. Adluru, H. Kang, C. Muralidharan, and V. Singh, "A digital mask to safeguard patient privacy," Nature Medicine, vol. 28, no. 9, pp. 1883–1892, 2022.

[2] M. El Helou, M. Mancini, and R. Timofte, "VerA: Versatile Anonymization Applicable to Clinical Facial Photographs," *arXiv preprint arXiv:2312.09976*, 2023.

[3] M. Rempe, S. Yao, and H. Ling, "De-Identification of Medical Imaging Data: A Comprehensive Tool for Ensuring Patient Privacy," *arXiv preprint arXiv:2410.12402*, 2024.

[4] R. Leyva, B. Meden, P. Peer, and V. Štruc, "Demographic bias effects on face image synthesis," In *Proceedings of CVPR*, 2024.

[5] M. Huber et al., "Bias and diversity in synthetic-based face recognition," In *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

[6] A. Bissoto et al., "Skin lesion synthesis with generative adversarial networks," In *Proceedings of OR 2.0 Workshop, MICCAI*, pp. 294–302, 2018.

[7] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," Medical Image Analysis, vol. 58, p. 101552, 2019.

[8] R. Zhang et al., "FairSkin: Fair Diffusion for Skin Disease Image Generation," *arXiv preprint arXiv:2410.22551*, 2024.

[9] M. A. Farooq et al., "Derm-t2im: Harnessing synthetic skin lesion data via stable diffusion models for enhanced skin disease classification using vit and cnn," In *Proceedings of EMBC*, 2024.

[10] Z. Liu, P. Luo, X. Wang, and X. Tang, "Large-scale celebfaces attributes (celeba) dataset," In *Proceedings of ICCV*, pp. 11–15, 2018.

[11] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," In *Proceedings of CVPR*, pp. 4401–4410, 2018

[12] N. Munia and A.-A.-Z. Imran, "DermDiff: Generative Diffusion Model for Mitigating Racial Biases in Dermatology Diagnosis," *arXiv preprint arXiv:2503.17536*, 2025.

[13] N. Munia and A.-A.-Z. Imran, "Prompting Medical Vision-Language Models to Mitigate Diagnosis Bias by Generating Realistic Dermoscopic Images," *arXiv preprint arXiv:2504.01838*, 2025.

[14] R. Rombach et al., "High-resolution image synthesis with latent diffusion models," In *Proceedings of CVPR*, pp. 10684–10695, 2022.

[15] J. Li et al., "Self-supervised Preference Optimization: Enhance Your Language Model with Preference Degree Awareness," In *Proceedings of EMNLP*, pp. 14452–14466, 2024.

[16] K. Lander, V. Bruce, and H. Hill, "Evaluating the effectiveness of pixelation and blurring on masking the identity of familiar faces," Applied Cognitive Psychology, vol. 15, no. 7, pp. 101–116, 2001.

[17] J. Todt, S. Hanisch, and T. Strufe, "Fantômas: Understanding Face Anonymization Reversibility," *arXiv preprint arXiv:2210.10651*, 2022.

[18] A. Ghorbani et al., "Dermgan: Synthetic generation of clinical skin images with pathology," In *Proceedings of PMLR*, pp. 155–170, 2020.

[19] S. I. Cho et al., "Generation of a melanoma and nevus data set from unstandardized clinical photographs on the internet," JAMA Dermatology, vol. 159, no. 11, pp. 1223–1231, 2023.

[20] P. Sakunchotpanit et al., "Representations of skin tone and sex in dermatology by generative artificial intelligence: a comparative study," Clinical and Experimental Dermatology, 2025.

[21] S. I. Ktena et al., "Generative models improve fairness of medical classifiers under distribution shifts," Nature Medicine, 2024.

[22] L. Sagers et al., "Improving dermatology classifiers across populations using images generated by large diffusion models," In *Proceedings of the NeurIPS 2022 Workshop*, 2022.

[23] D. Carrion and M. Norouzi, "FEDD – Fair, Efficient, and Diverse Diffusion-based Lesion Segmentation and Malignancy Classification," *In Proceedings of MICCAI*, 2023.

[24] A. Kirillov, Y. Wu, K. He, and R. Girshick, "Pointrend: Image segmentation as rendering," In *Proceedings of CVPR*, pp. 9799–9808, 2020.

[25] Rafael Rafailov, Abhishek Sharma, Mitchell Eisner, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, volume 36, 2024.

[26] X. Shen et al., "Finetuning Text-to-Image Diffusion Models for Fairness," In *Proceedings ICLR*, 2024.

[27] H. Hukkelås, R. Mester, and F. Lindseth, "Deepprivacy: A generative adversarial network for face anonymization," In *Proceedings of ISVC*, pp. 565–578, 2019.

[28] M. Le et al., "Anonfaces: Anonymizing faces adjusted to constraints on efficacy and security," In *Proceedings of WPES*, pp. 87–100, 2020.

[29] G. Huang et al., "Densely connected convolutional networks," In *Proceedings of CVPR*, pp. 4700–4708, 2017.

[30] InsightFace Team, "InsightFace: 2D and 3D Face Analysis Project," https://insightface.ai/, Accessed: 2023-09-03.