Controllable Generation of Diverse Dermatological Imagery for Fair and Efficient Malignancy Classification

Héctor Carrión[1*], Narges Norouzi[2]

University of California, Santa Cruz[1], University of California, Berkeley[2],
hcarrion@ucsc.edu*

Abstract. Accurate dermatological diagnosis naturally necessitates equitable performance across diverse populations, yet a systematic lack of expertly annotated images, especially for underrepresented skin tones and rare diseases, impedes progress toward measurably fair methods. We introduce **cgDDI** (**C**ontrollable **G**eneration of **D**iverse **D**ermatological **I**magery), a hybrid framework that (1) synthesizes realistic healthy skin samples without disturbing other input properties, (2) maps single-sample rare lesions onto novel skin-tones and locations non-parametrically, and (3) allows for efficient parametric generation with as few as 10 training samples. The framework supports both human and automated segmentation masking, enabling scalability to datasets without pre-made lesion masks. We grow a 656-image dataset by more than 400× and validate across two datasets: biopsy-confirmed Diverse Dermatology Images (DDI) and expert-verified Fitzpatrick17k (F17k). On the DDI benchmark, we achieve malignancy classification accuracy of 86.4% under synthetic-only training and 90.9% state-of-the-art performance with real data fine-tuning, alongside leading fairness metrics. Cross-dataset experiments show +13.9% accuracy improvements on unseen F17k data despite minimal disease overlap. We openly release 266k+ synthetic images, code, and generative models to further support fairness research at https://github.com/hectorcarrion/ControllableGenDDI.

Keywords: Dermatology · Fairness · Synthetic Data · Diffusion Models

1 Introduction

Skin diseases affect millions globally, with expert diagnosis accuracy being measurably lower for darker-skinned populations, even if the physicians originate from diverse backgrounds [14]. Furthermore, over 3 billion people lack access to adequate dermatological care, especially those in impoverished communities [7]. Early detection of skin cancers significantly increases survival rates [3], yet Artificial Intelligence (AI) systems trained on biased data sources risk exacerbating disparities. A survey of 70 dermatological AI studies found fewer than 25% included ethnicity and only 10% included skin-tone descriptors [9]. Recent generative approaches addressing this issue require large or private training

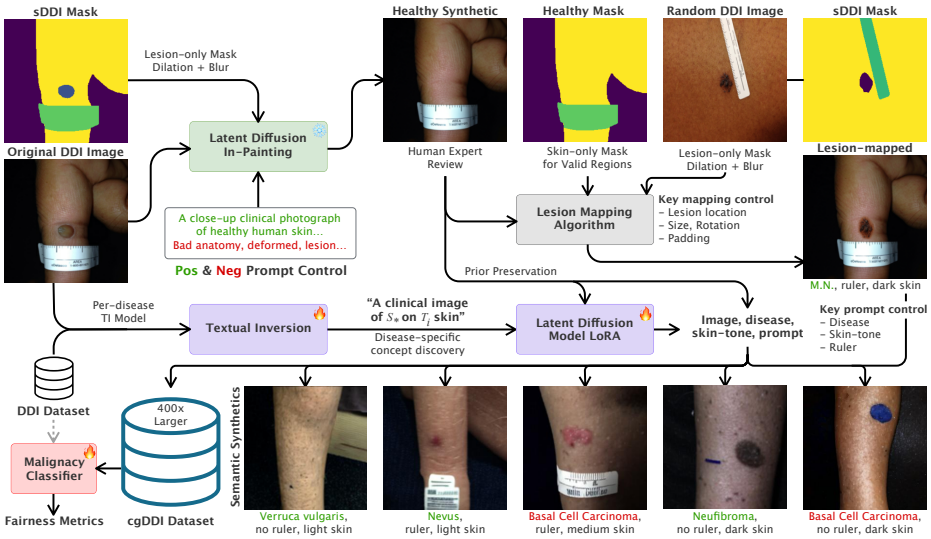

Fig. 1. cgDDI framework. Original images, masks, and prompts produce **healthy synthetics**. These serve as targets for **lesion-mapped synthetics**, as prior-preservation anchors, and as semantic prompts. Disease-specific concepts are learned via textual inversion and used to fine-tune latent diffusion models from which **semantic synthetics** are sampled. The aggregated data trains **fair classifiers**.

sets [2,18], do not cover the full skin-tone spectrum [25], or ignore extremely rare diseases [23]. We present cgDDI, a novel hybrid generation framework (Fig. 1) addressing these limitations through complementary parametric and non-parametric approaches. Our contributions are:

1. **cgDDI Framework:** A controllable method combining (a) latent diffusion inpainting for pixel-perfect healthy skin synthesis, (b) non-parametric lesion mapping enabling *single-sample* disease augmentation where parametric methods struggle, and (c) efficient parametric generation via textual inversion and Low Rank Adaptation (LoRA) regularized by, to our knowledge, the first use of prior-preservation loss (PPL) anchored on healthy images.
2. **cgDDI Dataset:** 266,136 skin-tone-balanced synthetic images (healthy, lesion-mapped, semantic) with fairness labels, openly released, including 309 pixel-perfect in-clinical-distribution healthy synthetics, a resource not present in prior work (Table 1).
3. **Classification and Fairness:** Malignancy classification on the DDI benchmark reaches SOTA accuracy (90.9%, up 3.5%) under full Fitzpatrick I–VI coverage including rich fairness metrics for both discriminative and generative results. Cross-dataset validation on F17k demonstrates generalizability with +4.7% intra- and +13.9% cross-dataset accuracy improvements.

Table 1. Comparison of synthetic dermatological datasets across training data, fairness indicators, skin-tone coverage, number of learned diseases, single-view augmentation support and whether the resulting datasets are published openly.

Method	Training Data	Fairness Metrics	FST Cov.	Ctrl. Diseases	Total Size	Single-view	Open Data
Sagers et al. (2022) [23]	F17k	FST Acc.	I–VI	3	192	–	–
Sagers et al. (2023) [22]	F17k, DDI	FST Acc.	I–VI	9	459k	–	✓
Akrout et al. (2024) [2]	Private	None	None	6	180k	–	–
Ktena et al. (2024) [18]	Private	FST Gap	I–VI	27	50k	–	–
Wang et al. (2024) [25]	F17k	FST Acc.	I–II, V–VI	7	7.6k	–	–
cgDDI (ours)	DDI	Multiple	I–VI	13	266k	✓	✓

2 Related Work

The F17k [15] dataset is widely used toward fairness research thanks to its large 17,000 sample size, however, it suffers from 3.6× skin-tone imbalance and $> 30\%$ label noise from non-expert annotations [8]. Recently an expert-verified F17k subset [14] has been released but it is much smaller in scope (364 samples). The DDI dataset [10] is relatively larger (656 samples) fully biopsy-confirmed with dermatologist-verified Fitzpatrick-scale labels, and mostly balanced across skin tones, with sDDI [6] providing segmentation masks. For these reasons, we train our main models on DDI but evaluate cross-dataset with expert-verified F17k. Recent malignancy classification fairness work has advanced through contrastive disentanglement [12] and patch alignment [1], establishing the DDI benchmark for both accuracy and fairness metrics which we compare against.

Diffusion models (DMs) [11] pre-trained for text-conditioned generation have shown controllability in dermatology [22], with textual inversion [13] improving control [25]. However, existing frameworks do not cover rare disease conditions with low frequencies, may rely on noisy or private training data, may not release generated outputs, several have incomplete skin-tone coverage and none report rich fairness metrics. We survey recent synthetic datasets in Table 1. Our work addresses each of these gaps.

3 cgDDI Framework

cgDDI generates three complementary types of synthetic dermatological imagery through a sequential pipeline. We first consolidate DDI (Sec. 3.1), develop a latent diffusion inpainting algorithm to create healthy samples (Sec. 3.2), which then serve as canvas for non-parametric lesion mapping (Sec. 3.3) and as prior-preservation anchors for parametric semantic generation (Sec. 3.4). We generate 309 healthy, 80,427 lesion-mapped, and 185,400 semantic synthetic images.

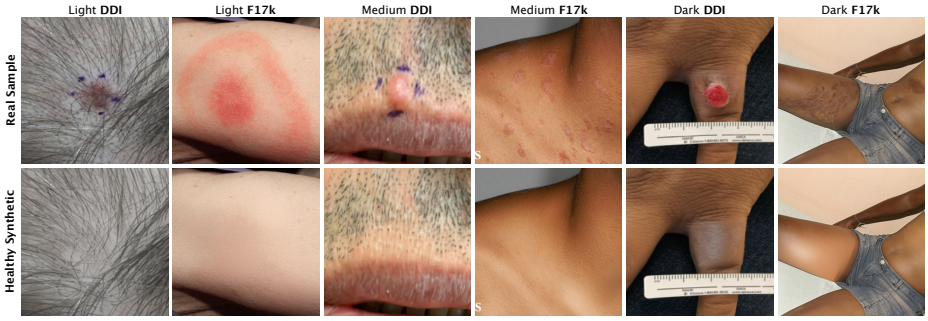

Fig. 2. Healthy synthetic imagery. Our inpainting removes target lesions (and markers), producing lesion-less reconstruction robust to hair, morphology, and body location. These samples are later used for lesion mapping and in-distribution prior preservation. Results shown for DDI (human masking) and F17k (algorithmic masking).

3.1 Data Pre-processing

DDI contains 656 samples (171 malignant, 485 benign). We consolidate 78 original disease labels into 65 categories based on histopathological similarity as in previous work [24], yielding 25 single-observation diseases, 27 diseases between 2 and 10 observations, and 13 diseases with >10 observations.

3.2 Healthy Synthesis via Latent Diffusion Inpainting

To our knowledge, a dataset providing dermatologist-verified healthy skin imagery collected analogously to diseased samples does not exist. We create it by inpainting lesion regions using a UNet denoiser (1.22B parameters) and MoVQ-GAN decoder [19], guided by positive (*"healthy, smooth, normal human skin"*) and negative (*"lesion, hole, transparent, eye"*) prompts. Segmentation masks delineate the inpainting regions; we apply dilation and Gaussian blur to smooth mask boundaries. Given 334 masked sDDI [6] inputs, we retain 309 healthy synthetics after human review (7% discard rate for generative artifacts). Results are shown in Fig. 2.

While we leverage human-made masks for DDI, our framework is compatible with algorithmic masking. We demonstrate this by masking F17k [14] via SAMv3 [5] for fully automated mask generation. SAMv3 successfully segments lesions across skin tones without dataset-specific training which our framework inputs directly, confirming generalizability and scaling to datasets without premade annotations. This is shown on Fig. 2. We note that SAMv3 masks can be noisy for medical images, and thus meet our discard criteria more often than human-made masks. Failure-mode examples are discussed on our repository.

3.3 Non-Parametric Lesion Mapping

For extremely rare diseases where parametric learning is infeasible, we transplant real lesions onto healthy canvases. Our algorithm leverages segmentation masks

to identify valid skin regions, and follows padding constraints (e.g. minimum 10 pixels from skin edges) to avoid placing lesions on unrealistic positions. Given 309 healthy images and 334 donor masks, we generate 80,427 lesion-mapped samples (22% discarded by padding constraints). This non-parametric approach enables augmentation from single-sample diseases zero-shot.

3.4 Parametric Semantic Generation

For diseases with $\geq$10 samples, we learn disease-specific tokens through textual inversion [13], then fine-tune the latent DM backbone [20] via LoRA [17]. Critically, we are the first in dermatological generation to employ Prior Preservation Loss (PPL) [21], using our healthy synthetics as an in-distribution regularization set. PPL addresses two issues in DM fine-tuning: semantic drift (forgetting class-level knowledge) and reduced output diversity, enabling more faithful generation than textual inversion alone [26]. Our healthy synthetics are uniquely suited for this role as they share the clinical imaging conditions and verified skin-tone labels of the training data.

At generation time, for healthy images $H{=}\{h_i\}_{i=1}^{309}$, diseases $D{=}\{d_j\}_{j=1}^{40}$, and skin tones $S{=}\{s_k\}_{k=1}^{3}$, each triple (h_i, d_j, s_k) produces $R{=}5$ samples:

$$x_{i,j,k}^{(r)} = f_{\theta,j}\big(h_i,\ \textit{“An image of } S_{*,j} \textit{ on a } s_k\textit{-toned individual”};\ \alpha,\ \beta,\ t\big) \quad (1)$$

where $f_{\theta,j}$ is the disease-specific DM, $S_{*,j}$ is the learned token, α is the conditioning strength, β the guidance scale, and t the inference steps. This yields $|H|{\times}|D|{\times}|S|{\times}R = 185{,}400$ semantic synthetics, balanced at 1,545 per skin tone per disease. We find $\sim$10 samples sufficient for viable generation quality, but release all synthesized images in our dataset repository for further study.

4 Experiments

4.1 Setup and Metrics

We adopt the PatchAlign [1] classifier (ViT-B/16) and evaluation protocol for direct comparison with prior state-of-the-art including their five-fold cross-validation with the same seeds. The PatchAlign training pipeline applies standard augmentations (crops, rotations, color jitter and flips); cgDDI gains are reported *on top of* this stack and operate orthogonally along attributes such as skin tone and lesion morphology. We report three fairness metrics: **PQD** (Predictive Quality Disparity), the ratio of worst-to-best skin-tone accuracy; **DPM** (Demographic Parity), measuring positive-prediction-rate consistency; and **EOM** (Equality of Opportunity), measuring true-positive-rate consistency and identified as the most important metric by [1]. Please see PatchAlign [1] for full formula definitions. We additionally evaluate generative quality via FID [16], KID [4], and LPIPS [27] between cgDDI and held-out real images, stratified by skin tone. Test-set details per dataset: DDI includes 131 test samples per fold (42 Light I–II, 48 Medium III–IV, 41 Dark V–VI); F17k 15% hold-out includes 55 test-set

Fig. 3. cgDDI samples. Row 1: real images used as prompts. Row 2: lesion from a donor transplanted onto the prompt image. Rows 3–4: semantic synthetics conditioned on the prompt, a target disease (malignant/benign), and target skin tone (light, medium, dark).

samples (22 Light, 17 Medium, 16 Dark). We recognize that in ideal conditions these test sets should be larger, and encourage the community to collect more sets of biopsy confirmed observations alongside accurate skin-tone labels.

4.2 Malignancy Classification

We run two experiments: *Exp. 1* trains purely on cgDDI synthetics; *Exp. 2* trains on synthetics then fine-tunes on real DDI data following [1]. Both are evaluated on held-out real DDI test sets with leakage prevention (i.e. excluding training synthetics conditioned on downstream test images).

Results are shown in Table 2. Exp. 1 achieves competitive accuracy (86.4%) while substantially improving all fairness metrics over prior methods. Exp. 2 achieves SOTA accuracy (90.9%) across all skin tones. EOM raises from 69.6 to 86.6, improving fairness significantly. Stratifying by disease rarity (Table 3), synthetic-only performance remains competitive, including under very rare conditions (1–2 samples: 83.3%), supporting lesion mapping for single-sample augmentation. Fine-tuning on real data primarily benefits common diseases (>10 samples), consistent with the real-life scarcity of rare samples.

Table 2. DDI classification and fairness. R. = real, S. = synthetic. Bold = best, underline = second best. Our method achieves SOTA accuracy and leading EOM.

	Accuracy (%) ± Std				**Fairness** ± Std		
Method	Mean	Light	Med.	Dark	PQD	DPM	EOM
Baseline (R.)	82.4±1.5	83.3±1.0	74.6±5.7	89.7±2.2	77.0±1.9	<u>75.2</u>±13	58.7±4.3
FairDisCo [12]	83.8±0.4	88.6±0.1	71.7±2.2	92.0±2.8	78.0±4.5	72.8±12	63.7±3.5
PatchAlign [1]	<u>87.4</u>±1.2	<u>89.6</u>±2.6	80.3±5.7	<u>92.3</u>±1.3	86.9±6.1	74.9±12	69.6±1.7
Exp. 1 (S.)	86.4±1.0	88.9±1.5	<u>84.1</u>±2.6	86.0±1.8	**94.6**±3.1	**82.0**±9.7	<u>81.9</u>±2.8
Exp. 2 (S.+R.)	**90.9**±1.3	**93.3**±2.2	**86.4**±4.1	**93.0**±1.0	<u>92.5</u>±2.5	68.8±11	**86.6**±1.9

Table 3. Disease rarity performance. Test-set accuracy stratified by the number of original real data observations per disease.

Disease Rarity	Cases in test set	Exp. 1 (S.) Accuracy (%)	Exp. 2 (S.+R.) Accuracy (%)
Common (>10 samples)	107	85.05	91.59
Rare (3–10 samples)	19	94.74	89.47
Very rare (1–2 samples)	6	83.33	83.33

4.3 Cross-Dataset Validation

To demonstrate generalizability beyond DDI, we process the expert-verified F17k subset [14] (364 samples) through our full pipeline using SAMv3 automated masking, producing 46 healthy, 1,124 lesion-mapped, and 5,520 semantic synthetics after discard criteria.

Intra-Dataset (F17k → F17k). Table 4 (top) shows that training on F17k synthetics then fine-tuning on real data achieves 90.7% accuracy (+4.7% over baseline) with the highest PQD, confirming cgDDI effectiveness on a second dataset.

Cross-Dataset Transfer. We synthesize inter-dataset data by mapping lesions and generating semantics across DDI and F17k, producing 23,216 lesion-mapped and 46,050 semantic cross-dataset synthetics. Table 4 (bottom) shows transfer results. Most notably, DDI-trained synthetics improve F17k accuracy by +13.9% (from 60.5% to 74.4%) despite only one shared disease condition (cutaneous T-cell lymphoma). Aggregated mixed-dataset training achieves up to 93.0% on F17k and 86.7% on DDI, with synthetics consistently improving over baselines.

4.4 Synthesis Ablation Study

We evaluate the individual contribution of each generation method by training classifiers on different subsets of cgDDI (Table 5). Training solely on healthy and lesion-mapped synthetics slightly reduces overall accuracy compared to real DDI (−1.2%) but improves medium skin-tone performance, likely due to the limited morphological diversity in non-parametric outputs. Semantic synthetics alone

Table 4. Cross-dataset classification results. Intra-dataset F17k validation (top), cross-dataset transfer (middle), and aggregated mixed-dataset training (bottom). Baselines are trained on real data only for the corresponding setting.

Setting	Method	Acc	Light	Med.	Dark	PQD	DPM	EOM
F17k → F17k	Baseline	86.0	86.7	82.4	**90.9**	0.906	**0.455**	0.500
	Synth Only	88.4	86.7	**94.1**	81.8	0.869	0.441	0.500
	Synth + Real	**90.7**	86.7	**94.1**	**90.9**	**0.921**	0.441	0.500
F17k → DDI	Baseline	**79.6**	**64.3**	**82.2**	**87.5**	0.735	0.500	0.500
	Synth Only	74.3	57.1	77.8	82.5	0.693	0.604	0.440
	Synth + Real	75.2	**64.3**	80.0	77.5	**0.804**	**0.678**	**0.703**
DDI → F17k	Baseline	60.5	66.7	47.1	72.7	0.647	0.515	0.526
	Synth Only	**74.4**	**80.0**	**70.6**	72.7	**0.882**	0.556	**0.613**
	Synth + Real	69.8	73.3	76.5	54.5	0.713	**0.664**	0.388
Mix → F17k	Baseline	86.0	86.7	88.2	81.8	**0.927**	**0.471**	0.500
	Synth Only	86.1	**93.3**	88.2	72.7	0.779	0.378	**0.750**
	Synth + Real	**93.0**	**93.3**	88.2	**100.0**	0.882	0.378	0.500
Mix → DDI	Baseline	83.2	75.0	82.2	90.0	0.833	**0.679**	0.784
	Synth Only	79.7	71.4	84.4	80.0	0.846	0.436	**0.833**
	Synth + Real	**86.7**	**82.1**	**84.4**	**92.5**	**0.888**	0.600	0.750

Table 5. Classification accuracy per data type.

Training Data	Mean	Light	Med.	Dark
Real DDI only	82.4	83.3	74.6	89.7
Healthy + Map	81.2	82.1	78.4	82.9
Semantic only	84.7	86.3	82.5	85.2
All cgDDI (Exp. 1)	**86.4**	**88.9**	**84.1**	86.0

Table 6. Generative quality per skin tone.

Skin Tone	FID ↓	KID ↓	LPIPS ↓
Light	103.45	0.039	0.715
Medium	88.41	0.032	0.741
Dark	108.07	0.016	0.734
Max/Min	1.22	2.41	1.04

provide a strong boost (+2.3% over real data), driven by parametric learning of disease-specific features. The combination of all three methods yields the best performance, validating our multi-pronged approach where each method addresses different scarcity scenarios.

4.5 Generative Fairness

We evaluate whether generation quality is equitable across skin tones via FID, KID, and LPIPS between cgDDI and held-out DDI images (Table 6). All metrics demonstrate reasonable stability: FID dispersion is low (σ=8.39, max/min ratio 1.22), LPIPS is near-identical across tones (σ=0.011). Each metric slightly favors a different tone (FID: Medium, KID: Dark, LPIPS: Light), indicating no systematic advantage for any population.

5 Conclusion

We present cgDDI, a hybrid generation framework that synthesizes fair and diverse dermatological imagery under extreme data constraints. By combining non-parametric lesion mapping with parametric generation and prior-preserving healthy priors, our method achieves state-of-the-art DDI classification (Accuracy 90.9%, up from 87.4% prior best [1]) and leading fairness metrics (EOM: 86.6%, up from 69.6%). Cross-dataset validation on F17k confirms generalizability, and compatibility with SAMv3 enables scalability without expert masks. We release 266k+ synthetic images, code, and models to support equitable dermatological AI research. Limitations include the minimum $\sim$10 samples needed for high-quality parametric generation (which our non-parametric processing addresses to a degree) and the benefit of incorporating board-certified dermatologist review for quality assessment. Future directions include few-shot adaptation techniques and cross-disease transfer learning to further reduce data requirements.

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

References

1. Aayushman, Gaddey, H., Mittal, V., Chawla, M., Gupta, G.R.: PatchAlign: Fair and Accurate Skin Disease Image Classification by Alignment with Clinical Labels. arXiv (Cornell University) (2024). https://doi.org/10.48550/arxiv.2409.04975
2. Akrout, M., Gyepesi, B., Holló, P., et al.: Diffusion-based data augmentation for skin disease classification: Impact across original medical datasets to fully synthetic images. In: Deep Generative Models. pp. 99–109. Springer Nature Switzerland (2024)
3. Balch, C.M., Gershenwald, J.E., Soong, S.J., et al.: Final version of 2009 AJCC Melanoma Staging and Classification. Journal of Clinical Oncology **27**(36), 6199–6206 (11 2009)
4. Binkowski, M., Sutherland, D.J., Arbel, M., et al.: Demystifying MMD GANs. International Conference on Learning Representations (2018)
5. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: Sam 3: Segment anything with concepts (2025)
6. Carrión, H., Norouzi, N.: FEDD – Fair, Efficient, and diverse diffusion-based lesion segmentation and malignancy Classification. arXiv (Cornell University) (2023). https://doi.org/10.48550/arxiv.2307.11654
7. Coustasse, A., Sarkar, R., Abodunde, B., et al.: Use of teledermatology to improve dermatological access in rural areas. Telemedicine and e-Health **25**, 1022–1032 (11 2019)

8. Daneshjou, R., Barata, C., Betz-Stablein, B., et al.: Checklist for evaluation of image-based artificial intelligence reports in dermatology. JAMA Dermatology **158**, 90 (01 2022)
9. Daneshjou, R., Smith, M.P., Sun, M.D., et al.: Lack of transparency and potential bias in artificial intelligence data sets and algorithms. JAMA Dermatology **157** (09 2021)
10. Daneshjou, R., Vodrahalli, K., Novoa, R.A., et al.: Disparities in dermatology ai performance on a diverse, curated clinical image set. Science Advances **8** (08 2022)
11. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. NIPS (06 2021)
12. Du, S., Hers, B., Bayasi, N., et al.: Fairdisco: Fairer ai in dermatology via disentanglement contrastive learning. Lecture Notes in Computer Science **13804**, 185–202 (2023)
13. Gal, R., Alaluf, Y., Atzmon, Y., et al.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: The Eleventh International Conference on Learning Representations (2023)
14. Groh, M., Badri, O., Daneshjou, R., et al.: Deep learning-aided decision support for diagnosis of skin disease across skin tones. Nature Medicine **30**(2), 573–583 (2 2024)
15. Groh, M., Harris, C., Soenksen, L., et al.: Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. CVPRW (04 2021)
16. Heusel, M., Ramsauer, H., Unterthiner, T., et al.: GANs trained by a two Time-Scale update rule converge to a local Nash equilibrium. Proceedings of the 31st International Conference on Neural Information Processing Systems **30**, 6626–6637 (2017), https://arxiv.org/pdf/1706.08500
17. Hu, E.J., Shen, Y., Wallis, P., et al.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
18. Ktena, I., Wiles, O., Albuquerque, I., et al.: Generative models improve fairness of medical classifiers under distribution shifts. Nature Medicine **30**(4), 1166–1173 (4 2024)
19. Razzhigaev, A., Shakhmatov, A., Maltseva, A., Arkhipkin, V., Pavlov, I., Ryabov, I., Kuts, A., Panchenko, A., Kuznetsov, A., Dimitrov, D.: Kandinsky: an Improved Text-to-Image Synthesis with Image Prior and Latent Diffusion. arXiv (Cornell University) (2023). https://doi.org/10.48550/arxiv.2310.03502
20. Rombach, R., Blattmann, A., Lorenz, D., et al.: High-Resolution Image Synthesis with Latent Diffusion Models. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 10674–10685 (6 2022)
21. Ruiz, N., Li, Y., Jampani, V., et al.: DreamBooth: Fine Tuning Text-to-Image diffusion models for Subject-Driven Generation. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 22500–22510 (6 2023)
22. Sagers, L.W., Diao, J.A., Melas-Kyriazi, L., Groh, M., Rajpurkar, P., Adamson, A.S., Rotemberg, V., Daneshjou, R., Manrai, A.K.: Augmenting medical image classifiers with synthetic data from latent diffusion models. arXiv (Cornell University) (2023). https://doi.org/10.48550/arxiv.2308.12453
23. Sagers, L.W., Diao, J.A., Groh, M., et al.: Improving dermatology classifiers across populations using images generated by large diffusion models. In: NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research (2022)
24. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data **5** (08 2018)

25. Wang, J., Chung, Y., Ding, Z., Hamm, J.: From Majority to Minority: A diffusion-based augmentation for underrepresented groups in skin lesion analysis. arXiv (Cornell University) (2024). https://doi.org/10.48550/arxiv.2406.18375
26. Zeng, Y., Suganuma, M., Okatani, T.: An improved method for personalizing diffusion models. arXiv (Cornell University) (2024). https://doi.org/10.48550/arxiv.2407.05312
27. Zhang, R., Isola, P., Efros, A.A., et al.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)