DiDGen: Diffusion-based Dual-task Synthesis for Dermoscopic Data Generation

Junjie Shentu, Matthew Watson, and Noura Al Moubayed*

Durham University, Durham DH1 3LE, UK
*corresponding author: noura.al-moubayed@durham.ac.uk

Abstract. Computer-aided diagnosis (CAD) systems for skin lesion analysis reduce costs and workload associated with the manual inspection of skin diseases. Nevertheless, the performance of deep learning (DL)-based CAD systems is constrained by the limited availability of labeled data, necessitating advanced dataset augmentation techniques. To address this limitation, we propose **DiDGen**, a novel method that employs **Di**ffusion models (DMs) for **D**ermoscopic image **Gen**eration and lesion-mask pair synthesis. Specifically, we introduce DermPrompt, a new type of structured text prompt rich with clinical details annotated by large language models (LLMs), which facilitates DMs' learning of fine-grained visual representations. Additionally, we propose a new paradigm for lesion-mask pair synthesis by incorporating a region-aware attention loss during finetuning to facilitate the build of semantic connections between text and visual representations, and then integrating test-time layout guidance with attention-based annotation to synthesize diverse and accurate lesion-mask pairs in a training-free manner. Extensive experiments demonstrate that our method improves the quality and diagnostic utility of generated dermoscopic images, thereby enhancing DL model performance in skin lesion classification and segmentation tasks. Our code is available at https://github.com/junjie-shentu/DiDGen.

Keywords: Dermoscopy · Diffusion Model · Skin Lesion Diagnosis

1 Introduction

Skin cancer is a public health challenge, with the most lethal melanoma causing 8290 deaths in the US in 2024 [24]. While dermoscopy is a reliable imaging modality for skin cancer diagnosis, manual visual assessment is time-intensive and prone to subjective bias, causing a diagnostic accuracy of about 60% among dermatologists [19], and driving demand for computer-aided diagnosis (CAD) systems. Current CAD approaches for skin lesion diagnosis predominantly focus on lesion segmentation and classification [10], However, their data-driven nature is constrained by the scarcity of public dermoscopic data [19]. Recent advances in generative models provide promising solutions for medical image generation, enhancing both performance and equity in downstream CAD tasks [14].

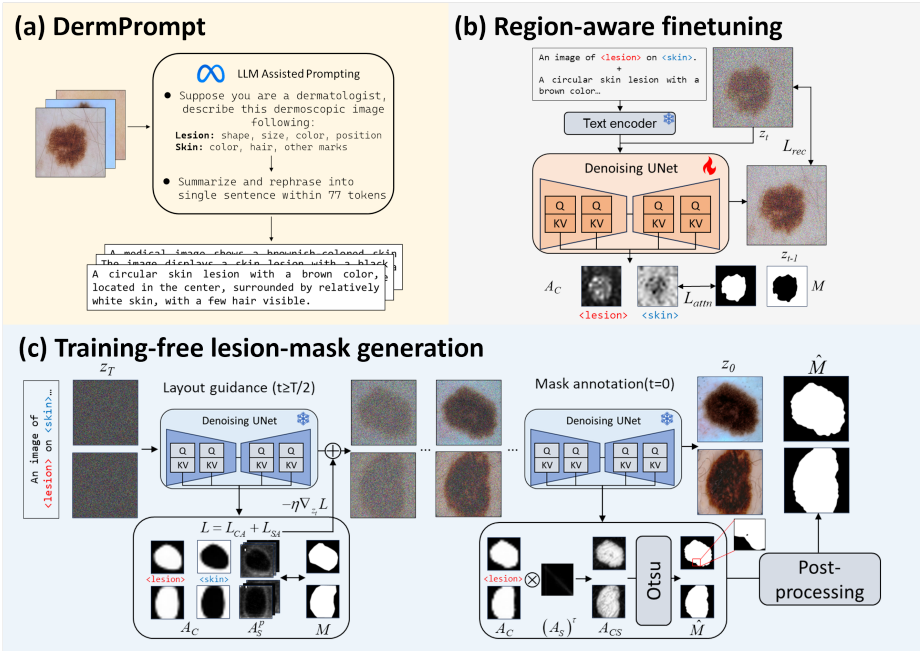

Fig. 1. Overview of our proposed method including three technical contributions

Generative adversarial network (GAN)-based methods have been proposed for dermoscopic image generation [20, 19, 2], yet struggle with fidelity and controllability. Diffusion models (DMs) [21] address these limitations through high-fidelity generation and text-guided control. While recent work finetunes DMs on dermoscopic data [23, 9], the use of simplistic text prompts underutilizes their semantic potential. Concurrently, lesion-mask generation frameworks designed to augment segmentation datasets using GANs [1] or DMs [7] require task-specific training, limiting multi-task applicability.

To overcome these limitations, we propose **DiDGen**, a novel and unified framework leveraging pretrained Stable Diffusion (SD) model for dual-task dermoscopic synthesis including dermoscopic images and lesion-mask pairs. First, we introduce DermPrompt, a novel type of structured text prompt enriched with clinical details annotated by large language models (LLMs), enabling fine-grained visual representation learning. Moreover, we propose a novel paradigm for image-mask pair synthesis that first builds semantic connections using an attention-based alignment strategy during finetuning, and then generates images paired with the corresponding masks using a training-free pipeline. Crucially, our method only requires finetuning the SD once to realize both generation tasks, and presents superior performance improvement for downstream tasks, offering an efficient solution for dermoscopic dataset augmentation.

2 Proposed Method

We begin by providing a brief overview of the SD model, and then introduce our method that includes several novel techniques, as illustrated in Fig. 1.

2.1 Preliminary

The SD model encodes an input image $x \in \mathbb{R}^{H\times W\times 3}$ into a latent representation $z \in \mathbb{R}^{h\times w\times c}$. Text prompts y are encoded into embeddings τ_θ using a pre-trained text encoder [21]. The UNet denoiser ε_θ is trained to predict the standard Gaussian noise ε given the corrupted latent z_t, the timestep t through:

$$L_{rec} = \mathbb{E}_{z,y,t,\varepsilon}\left[\|\varepsilon - \varepsilon_\theta\left(z_t, t, \tau_\theta(y)\right)\|_2^2\right] \tag{1}$$

The UNet integrates self-attention (SA) and cross-attention (CA) layers to capture the dependencies within the input data [21]. The SA layers capture the intra-image correspondence while the CA layers learn image-text interactions. The CA map A_C and SA map A_S can be calculated as follows:

$$A_C = softmax\left(Q_I K_T^\top/\sqrt{d}\right), A_S = softmax\left(Q_I K_I^\top/\sqrt{d}\right) \tag{2}$$

where Q_I, K_I, K_T are the query matrix, key matrix of z_t, and query matrix of $\tau_\theta(y)$, respectively. d denotes the latent dimension.

2.2 Attribute-aware DermPrompt

Instead of using simplistic and fixed text prompts for training and sampling [23, 9], we propose DermPrompt, which utilizes LLMs to generate structured text prompts that encapsulate clinical details. Specifically, we leverage Llama3 [8] with specially designed text prompts for visual captioning of dermoscopic images, and extract fine-grained attributes such as the shape, size, color, position of the lesion, skin color, and the presence of hair and other markings. Subsequently, we use Llama3 to further summarize and rephrase each image caption into a single sentence containing fewer than 77 tokens, thereby adapting to the input capacity of the SD model, as illustrated in Fig. 1(a).

Clinically detailed DermPrompt enhances semantic grounding via attribute-aware training with dermatology-related lexicons. It directs the SD model to learn fine-grained visual representations, thereby reducing ambiguity. Beyond training, DermPrompt is also employed during sampling, with Llama3 generating new prompts from the existing DermPrompt to further improve the fidelity, diversity, and granularity of generated images.

2.3 Region-aware Finetuning of the SD model

During finetuning, we prepend a prefix " *An image of <lesion> on <skin>.*" to the DermPrompt, where "*<lesion>*" and "*<skin>*" are special tokens initialized by text embeddings of tokens *lesion* and *skin*, referred to as Placeholder

Tokens (P-Tokens). Their embeddings remain fixed; however, the corresponding CA maps are extracted and used as input for the region-aware attention loss to establish semantic connections between visual representations and text tokens.

The pre-trained SD model presents robust semantic connections between text prompts and images through CA maps [3] after extensive pre-training. However, in domain adaption scenarios where only limited data are available, it is challenging to establish such robust semantic connections between text and dermoscopic images. To address this issue, we introduce a region-aware CA loss as follows:

$$L_{attn} = \frac{1}{2} \sum_{i \in \{l,s\}} \| A_C\left(v_i, z_t\right) - M_i \|_2^2 \tag{3}$$

where $A_C\left(v_i, z_t\right)$ is the CA map at a scale of 16×16 for token v_i [11]. l and s denote the P-Tokens "*<lesion>*" and "*<skin>*", and M_l, M_s are the corresponding masks. L_{attn} penalizes the model to build semantic connections between the P-Tokens and corresponded regions in dermoscopic images, thus termed region-aware attention loss. The pre-trained SD model jointly optimizes L_{rec} and L_{attn} with a scaling coefficient α (empirically set to 0.1), the overall loss is given by:

$$L = L_{rec} + \alpha L_{attn} \tag{4}$$

2.4 Training-free Pipeline for Lesion-mask Generation

Leveraging the semantic connections built by the region-aware attention loss, we propose a training-free pipeline that generates diverse and precise lesion-mask pairs by combining test-time layout guidance and attention-based annotation [28, 5, 27, 15]. A schematic diagram is shown in Fig. 1(c).

The sampling process of SD can be controlled by the classifier guidance technique to realize test-time conditional sampling [6]. Furthermore, classifier guidance can be extended to layout guidance by regularizing attention maps with masks M, thereby controlling the position, size, and shape of the target object [5, 28, 18]. Our method considers both CA and SA regularizations. Specifically, we extract the CA maps of P-Tokens and apply a Gaussian filter on them to smooth the attention activation [3]. We formulate the CA regularization analogously to Eq. (3), renaming the loss function from L_{attn} to L_{CA} in this context. L_{CA} guides the lesion's position, size, and shape to resemble the layout of the condition mask M. While CA maps reveal the semantic connections, SA maps capture pixel-wise visual correspondences in the latent feature map z_t. Therefore, regularizing SA maps further restricts the layout to that of M. For each highlighted pixel $\hat{p} \in M$, we extract the corresponding SA map A_S^p at a scale of 32×32 [15] of the pixel p in z_t with the same coordinate. We then regularize A_S^p to reduce attention connections outside the region specified by M. The regularization term L_{SA} is defined as [18]:

$$L_{SA} = \frac{1}{2} \sum_{i \in \{l,s\}} \sum_{p_i \in M_i} \frac{\sum A_S^{p_i,B}}{\sum (1 - M_i)}, A_S^{p,B} = A_S^p \cdot (1 - M) \tag{5}$$

where $A_S^{p,B}$ is the background of A_S^p. Furthermore, since L_{SA} can reduce extraneous attention connections outside the region defined by M, it is helpful for mask annotation(see the ablation study in Section 3.4). During sampling, layout guidance updates z_t by minimizing both losses via gradient descent:

$$\hat{z}_t \leftarrow z_t - \eta \bigtriangledown_{z_t} (L_{CA} + L_{SA}) \tag{6}$$

where η is the learning rate empirically set to 20. Because classifier guidance is most effective in the early sampling stage [5], we apply the guidance for the first 50% timesteps, where the optimized latent map $\hat{z}_t$ substitutes for z_t to compute the subsequent latent map z_{t-1}. By adjusting η and the guiding range, our method can generate numerous images from the same input mask.

Nevertheless, achieving a precise pixel-level correspondence between the input mask and the generated image remains challenging for test-time layout guidance [5]. To address this challenge, we propose a method for simultaneously generating lesion-mask pairs by leveraging CA and SA maps in the final sampling stage [27, 15]. Observing that SA maps can sharpen the boundary of CA maps [13], we derive the mask $\hat{M}$ for the generated lesion by multiplying the CA map $A_{C_{\hat{t}}}$ of the P-Token *<lesion>* with the SA map $A_{S_{\hat{t}}}$, followed by the application of Otsu's thresholding [17]:

$$A_{CS} = (A_{S_{\hat{t}}})^{\tau} \cdot A_{C_{\hat{t}}}, \hat{M} = Otsu\,(A_{CS}) \tag{7}$$

where τ is an exponent empirically set to 4 [15]. Both maps are extracted at the final sampling step (i.e., $\hat{t} = 0$) where semantic and spatial information are most abundant [16]. Finally, we post-process $\hat{M}$ first applying dilation followed by erosion to fill small holes in the masks, as shown in Fig. 1(c).

3 Experiments

We perform three experiments to evaluate the general quality of generated images and their effect on classification, as well as the effect of generated lesion-mask pairs on segmentation. We utilize the pre-trained SD v2.1 as the backbone, and finetune it by 20000 steps with a learning rate of 1×10^{-5} and a batch size of 4. All experiments are conducted on an NVIDIA A100 GPU. We leverage the ISIC 2018 dataset [4], which provides 2594 lesion-mask pairs (Task 1) and 10015 dermoscopic images across seven diagnostic classes (Task 3), enabling multi-task training and evaluation including multi-class classification and segmentation.

3.1 General generation quality

We perform the general generation fidelity using Fréchet Inception Distance (FID) and Multi-scale Structural Similarity Index (MS-SSIM) between 1000 synthetic and real images from Task 1's test set. Diversity is measured via Learned Perceptual Image Patch Similarity (LPIPS) within the 1000 generated images. Comparative baselines include SL-StyleGAN [19], PGAN [12], DreamBoooth

Table 1. General genration quality of different models

	SL-StyleGAN	PGAN	Finetune	DreamBooth	Ours-prefix	Ours-DermPrompt
FID ↓	9.69	86.17	12.76	29.04	9.26	**8.96**
MS-SSIM ↑	0.34	0.37	0.39	**0.40**	0.38	0.39
LPIPS ↑	0.56	0.36	0.55	0.48	0.56	**0.58**

[22], and standard SD finetuning, all trained on Task 1 data. Notably, we employ text-only conditioning in this experiment, and apply the prefix as prompt for DreamBoooth and standard finetuning. As shown in Table 1, our method with DermPrompt achieves state-of-the-art FID and LPIPS scores, with near-optimal MS-SSIM score. Remarkably, even when using only the prefix as the prompt during sampling, our approach maintains competitive performance, demonstrating robust representation learned with DermPrompt.

3.2 Dataset Augmentation for Multi-class Classification

To assess the capacity of generative models for capturing clinically relevant features, we evaluate their ability to generate images across distinct diagnostic categories. We train our proposed method and the baselines introduced in Section 3.1 on Task 3 data while excluding the overrepresented "NV" class (66.9% prevalence) to mitigate bias. For the remaining six categories (MEL, BCC, AKIEC, BKL, DF, VASC), we generate 1000 samples per class. Figure 2(a) presents samples generated by different models. The results indicate that GAN-based models exhibit limited fidelity and insufficient representation of diagnostic characteristics, whereas diffusion-based models demonstrate superior fidelity and effectively capture diagnostic characteristics.

Subsequently, we train three classifiers, VGG16, DenseNet121, and ViT, using both the original training set and an augmented training set. Specifically, the augmented set comprises the original set and 6000 synthesized images (1000 per category). The original validation set is employed for early stopping during training, with a patience of 10 epochs. Classifier performance is evaluated on the testing set using micro precision, recall, and F1 score, as shown in Table 2. The results indicate that when the dataset is augmented by our method, DenseNet121 and ViT achieve optimal performance. We also perform paired t-tests on the classification results from classifiers trained on the original set and augmented set by our method. The calculated p-value is 1.35×10^{-3}, indicating the robustness of our method.

3.3 Dataset Augmentation for Segmentation

We evaluate the quality of lesion-mask pairs by training two segmentation models, DCSAU-Net [29] and XB-Former [25], on the original training set of Task 1 and augmented sets, using the original validation set for early stopping with a patience of 20 epochs. We compare our method with training-based image-translation models (Pix2PixHD [26], ControlNet [30]), and training-free layout

Table 2. Performance of classifiers trained on original and augmented datasets. Note that DSNet, DP are abbreviations of DenseNet121, and DermPrompt, respectively.

Model	Precision ↑			Recall ↑			F1 Score ↑		
	VGG16	DSNet121	ViT	VGG16	DSNet121	ViT	VGG16	DSNet121	ViT
Original set	0.782	0.811	0.834	0.775	0.813	0.834	0.775	0.811	0.830
SL-StyleGAN	0.790	0.801	0.831	0.792	0.804	0.832	0.788	0.804	0.832
PGAN	0.793	0.817	0.838	0.791	0.822	0.841	0.786	0.816	0.837
Finetune	**0.807**	0.823	0.840	**0.808**	0.822	0.843	**0.803**	0.819	0.840
DreamBooth	0.797	0.798	0.825	0.795	0.802	0.831	0.793	0.798	0.824
Ours-DP	0.805	**0.826**	**0.846**	0.805	**0.828**	**0.847**	0.801	**0.825**	**0.845**

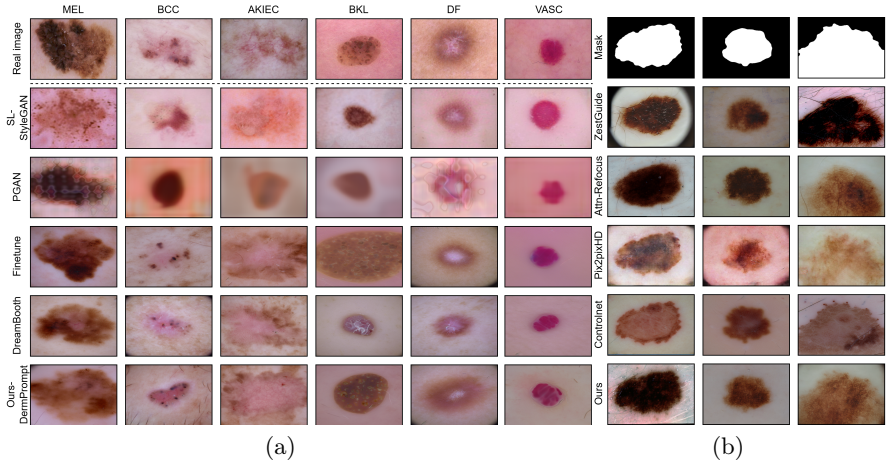

Fig. 2. Qualitative comparison of different models in (a) generation of dermoscopic images with different diagnoses and (b) generation of lesion-mask pairs

guidance methods (ZestGuide [5], Attn-Refocus [18]). To avoid information leakage, we first apply our method to generate 2500 lesion-mask pairs guided by masks from the training set. The generated masks retain the core structure of the original training masks but introduce minor differences, ensuring they can be safely used as inputs for baseline models without risking data leakage.

As shown in Fig. 2(b), training-free methods struggle to replicate precise mask layouts, resulting in minor discrepancies in the generated lesions. In contrast, training-based models capture layout details accurately and produce lesions with precise shapes, but they require significant computational resources for training. Our method generates highly consistent lesion-mask pairs without additional training, offering an efficient solution for augmenting segmentation datasets. Furthermore, we evaluate segmentation performance using the Dice coefficient (Dice) and Intersection over Union (IoU) on two scales, which are a small scale (1000 samples from the training set, S_{1k}) and a large scale (the full training set, $S_{2.5k}$), thereby demonstrating the effect of dataset augmentation under varying data scarcity. The results in Table 3 show that both segmen-

Table 3. Performance of segmentation models trained on original and augmented datasets. Note that the Red font denotes IoU and the Blue font denotes Dice.

Model	Training -free	DCSAU-Net			XB-Former		
		S_{1k}	$S_{1k}+2.5k$	$S_{2.5k}+2.5k$	S_{1k}	$S_{1k}+2.5k$	$S_{2.5k}+2.5k$
Pix2PixHD	×	0.723/	0.731/0.841	0.778/0.874	0.773/	0.790/0.881	0.819/0.890
ControlNet	×	0.836	0.742/0.848	0.775/0.871	0.868	**0.796**/**0.884**	0.823/0.891
ZestGuide	√	$S_{2.5k}$	0.727/0.839	0.764/0.865	$S_{2.5k}$	0.776/0.872	0.816/0.883
Attn-Refocus	√	0.768/	0.740/0.847	0.772/0.869	0.810/	0.784/0.876	0.811/0.885
Ours	√	0.867	**0.756**/**0.858**	**0.788**/**0.880**	0.884	0.788/0.880	**0.826**/**0.895**

tation models exhibit improved performance on the augmented datasets. Our dataset augmentation method enables DCSAU-Net to achieve the best performance at both scales, and allows XB-Former to obtain the best performance at the large scale and the second-best performance at the small scale, with a marginal performance gap compared to ControlNet and results comparable to Pix2PixHD. Furthermore, the p-value from paired t-tests on the segmentation results is 6.64×10^{-29}, reflecting the robust gain benefited from our method.

3.4 Ablation Study

First, we verify the region-aware attention loss L_{attn} by comparing CA maps of the P-Tokens, finding that CA maps can delineate lesion and skin regions when the model is finetuned with L_{attn} (Fig. 3(a)). Furthermore, we clarify that the CA guidance L_{CA} can increase the generation diversity, as evidenced by a higher LPIPS score for masks with L_{CA} (0.403) compared to those without L_{CA} (0.172). Finally, SA guidance L_{SA} can prevent mask thresholding failures. As shown in Fig. 3(b), applying L_{SA} corrects the lower values in A_{CS} map caused by artificial marks that may confuse Otsu's thresholding.

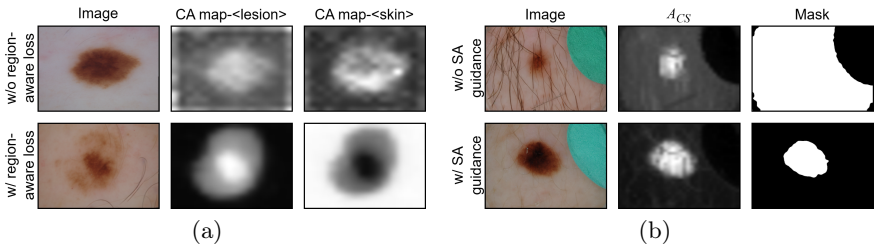

Fig. 3. Ablation studies on (a) region-aware attention loss and (b) SA guidance

4 Conclusion

We introduce **DiDGen**, a novel diffusion-based model for dermoscopic generation. Specifically, we propose DermPrompt, a structured text prompt paradigm

containing rich clinical details annotated by LLMs. Additionally, our approach generates accurate lesion-mask pairs by incorporating a region-aware attention loss during finetuning and integrating test-time layout guidance with attention-based mask annotation. With a single training run, DiDGen can produce both dermoscopic images and lesion-mask pairs, saving computational resources. Extensive experiments demonstrate that our model delivers superior performance improvements for downstream tasks compared to baseline models.

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

References

1. Abhishek, K., Hamarneh, G.: Mask2lesion: Mask-constrained adversarial skin lesion image synthesis. In: International workshop on simulation and synthesis in medical imaging. pp. 71–80. Springer (2019)
2. Bisla, D., Choromanska, A., Berman, R.S., Stein, J.A., Polsky, D.: Towards automated melanoma detection with deep learning: Data purification and augmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 0–0 (2019)
3. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM transactions on Graphics (TOG) **42**(4), 1–10 (2023)
4. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)
5. Couairon, G., Careil, M., Cord, M., Lathuiliere, S., Verbeek, J.: Zero-shot spatial layout conditioning for text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2174–2183 (2023)
6. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
7. Du, S., Wang, X., Lu, Y., Zhou, Y., Zhang, S., Yuille, A., Li, K., Zhou, Z.: Boosting dermatoscopic lesion segmentation via diffusion models with visual and textual prompts. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2024)
8. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
9. Farooq, M.A., Yao, W., Schukat, M., Little, M.A., Corcoran, P.: Derm-t2im: Harnessing synthetic skin lesion data via stable diffusion models for enhanced skin disease classification using vit and cnn. arXiv preprint arXiv:2401.05159 (2024)
10. Hasan, M.K., Ahamad, M.A., Yap, C.H., Yang, G.: A survey, review, and future trends of skin lesion segmentation and classification. Computers in Biology and Medicine **155**, 106624 (2023)
11. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: The Eleventh International Conference on Learning Representations (2023)

12. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
13. Khani, A., Taghanaki, S.A., Sanghi, A., Amiri, A.M., Hamarneh, G.: Slime: Segment like me. The Twelfth International Conference on Learning Representations (2024)
14. Ktena, I., Wiles, O., Albuquerque, I., Rebuffi, S.A., Tanno, R., Roy, A.G., Azizi, S., Belgrave, D., Kohli, P., Cemgil, T., et al.: Generative models improve fairness of medical classifiers under distribution shifts. Nature Medicine pp. 1–8 (2024)
15. Nguyen, Q., Vu, T., Tran, A., Nguyen, K.: Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. Advances in Neural Information Processing Systems **36**, 76872–76892 (2023)
16. Nguyen, T.T., Nguyen, D.A., Tran, A., Pham, C.: Flexedit: Flexible and controllable diffusion-based object-centric image editing. arXiv preprint arXiv:2403.18605 (2024)
17. Otsu, N., et al.: A threshold selection method from gray-level histograms. Automatica **11**(285-296), 23–27 (1975)
18. Phung, Q., Ge, S., Huang, J.B.: Grounded text-to-image synthesis with attention refocusing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7932–7942 (2024)
19. Qin, Z., Liu, Z., Zhu, P., Xue, Y.: A gan-based image synthesis method for skin lesion classification. Computer Methods and Programs in Biomedicine **195**, 105568 (2020)
20. Ren, Z., Guo, Y., Stella, X.Y., Whitney, D.: Improve image-based skin cancer diagnosis with generative self-supervised learning. In: 2021 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE). pp. 23–34. IEEE (2021)
21. 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)
22. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22500–22510 (2023)
23. Shavlokhova, V., Vollmer, A., Zouboulis, C.C., Vollmer, M., Wollborn, J., Lang, G., Kübler, A., Hartmann, S., Stoll, C., Roider, E., et al.: Finetuning of glide stable diffusion model for ai-based text-conditional image synthesis of dermoscopic images. Frontiers in medicine **10**, 1231436 (2023)
24. Siegel, R.L., Giaquinto, A.N., Jemal, A.: Cancer statistics, 2024. CA: a cancer journal for clinicians **74**(1), 12–49 (2024)
25. Wang, J., Chen, F., Ma, Y., Wang, L., Fei, Z., Shuai, J., Tang, X., Zhou, Q., Qin, J.: Xbound-former: Toward cross-scale boundary modeling in transformers. IEEE Transactions on Medical Imaging **42**(6), 1735–1745 (2023)
26. 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)
27. Wu, W., Zhao, Y., Chen, H., Gu, Y., Zhao, R., He, Y., Zhou, H., Shou, M.Z., Shen, C.: Datasetdm: Synthesizing data with perception annotations using diffusion models. Advances in Neural Information Processing Systems **36**, 54683–54695 (2023)

28. Xie, J., Li, Y., Huang, Y., Liu, H., Zhang, W., Zheng, Y., Shou, M.Z.: Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7452–7461 (2023)
29. Xu, Q., Ma, Z., Na, H., Duan, W.: Dcsau-net: A deeper and more compact split-attention u-net for medical image segmentation. Computers in Biology and Medicine **154**, 106626 (2023)
30. 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)