Domain-Adaptive Transformer GAN for High-Fidelity Skin Lesion Image Generation

Tauhid Hasan Department of Computer Science
American International University-Bangladesh
Dhaka, Bangladesh
22-46438-1@student.aiub.edu
Nura Alam Department of Computer Science
American International University-Bangladesh
Dhaka, Bangladesh
22-46457-1@student.aiub.edu
Mahmudus Sami Maahi Department of Computer Science
American International University-Bangladesh
Dhaka, Bangladesh
22-46446S-1@student.aiub.edu
Mostafijur Rahmman Department of Computer Science
American International University-Bangladesh
Dhaka, Bangladesh
22-47161-1@student.aiub.edu

Abstract

TransGAN, a transformer-based generative adversarial network originally developed for image synthesis, presents a compelling alternative to convolutional architectures by leveraging self-attention mechanisms to capture complex spatial dependencies. In this research, we adapt and optimize TransGAN for skin lesion image synthesis using the ISIC skin lesion dataset. To enhance generative performance, we introduce several architectural modifications: replacing instance pixel shuffle with CNN-based upsampling, substituting standard 2× average pooling with CNN average pooling, incorporating positional embeddings before each transformer block, and replacing the generator's final linear unflattening layer with a convolutional layer. These changes allow for a reduction in the number of attention heads per transformer block while maintaining image fidelity. The proposed architecture demonstrates improved synthesis quality, achieving a Fréchet Inception Distance (FID) score of 7.5, indicating high realism and diversity in the generated dermatological images. This approach contributes to more effective data augmentation for medical image analysis and training in dermatological applications.

Index Terms

GAN, Transformer, CNN, FID

I. Introduction

Generative Adversarial Networks (GANs) have achieved significant success across various computer vision tasks, including image synthesis, translation, and super-resolution [5], [6]. Despite these advances, GANs are known for their instability during training, which has prompted extensive research on stabilization techniques. These include regularization methods [7], [8], improved loss functions [5], [9], [10], and optimized training strategies [**?**], [**?**]. An emerging focus area has been on enhancing GAN architectures themselves. Studies such as [11], [7] have shown that across multiple datasets, widely used neural backbones often yield comparable performance, underscoring the need for more architecture-aware design improvements.

Among recent architectural advancements, TransGAN, introduced by Yifan Jiang et al. [12], proposed a convolution-free transformer-based GAN that replaces traditional convolutional operations with pure self-attention blocks. TransGAN demonstrated competitive results on natural image datasets, highlighting the potential of transformers in generative modeling. However, its application in medical imaging, where spatial detail and textural integrity are vital, remains underexplored.

One domain where generative modeling can provide substantial impact is melanoma detection. Melanoma is one of the deadliest forms of skin cancer, and early detection plays a crucial role in patient outcomes. However, collecting large, balanced datasets for training diagnostic models remains a significant challenge. To address this, synthetic image generation can help augment datasets and reduce annotation dependency. In this study, we utilize the Skin Lesion Towards Melanoma Detection dataset [13], a curated collection of dermoscopic images used in melanoma classification research.

Building on Jiang et al.'s TransGAN [12], we propose a modified architecture specifically adapted for the melanoma detection domain. Our architectural contributions include replacing instance pixel shuffle with CNN-based upsampling, using CNN-based average pooling instead of standard $2\times$ average pooling, inserting positional embeddings before each transformer block

to maintain spatial awareness, and replacing the generator's final linear unflattening layer with a convolutional layer. These modifications not only enhance image realism but also allow for a reduction in the number of attention heads per transformer block, making the model more efficient.

Using this architecture, we train and fine-tune our model on the Skin Lesion Towards Melanoma Detection dataset [13] and achieve a Fréchet Inception Distance (FID) score of 7.5, demonstrating the high quality and diversity of the generated skin lesion images.

This research contributes the following:

- **Modified Transformer-based GAN:** We propose an enhanced TransGAN architecture tailored for medical image synthesis, improving spatial coherence and reducing model complexity through CNN integration and positional encoding.
- **Domain-specific Adaptation and Training:** We train and fine-tune the proposed model on a melanoma-specific dataset, addressing the real-world challenge of limited and imbalanced dermatological data.
- **Quantitative and Qualitative Evaluation:** Our model achieves a strong FID score of 7.5, supporting its effectiveness in generating high-quality synthetic skin lesion images for potential use in training and augmentation pipelines in melanoma detection systems.

These contributions demonstrate the feasibility and potential of transformer-based GANs in medical image synthesis, particularly in data-scarce applications like melanoma classification.

II. Literature Review

The integration of transformer architectures into generative adversarial networks (GANs) has gained significant momentum in recent years, particularly in the domain of high-fidelity image synthesis. A number of contemporary models—namely, Styleformer, HiT, and StyleSwin—illustrate the diverse strategies and innovations that characterize this rapidly evolving research area.

Styleformer: Park et al. [1] introduced *Styleformer*, a transformer-based framework designed for latent space editing within GANs. The model incorporates latent masking within transformer blocks, facilitating selective manipulation of style-related features while preserving the semantic structure of the generated images. This masking mechanism enables targeted control in the latent domain, thereby achieving style-consistent generation. Empirical evaluations demonstrate Styleformer's robust performance across multiple benchmark datasets, achieving scores of 10.00 Inception Score (IS) on CIFAR-10, 3.92 Fréchet Inception Distance (FID) on CelebA, and 11.01 IS / 15.17 FID on STL-10, confirming its effectiveness in style preservation and structural coherence.

HiT: Building upon the premise of hierarchical representation learning, Zhao et al. [2] proposed *HiT* (Hierarchical Transformer), which adopts a patch-wise hierarchical transformer design. The model utilizes hierarchical latent masking to construct a multi-level feature representation, thereby capturing both global and local contextual information crucial for high-resolution image generation. Experimental results on the FFHQ and CelebA-HQ datasets yield FID scores of 2.95 and 3.39, respectively, underscoring HiT's capability to generate photo-realistic facial images with fine-grained detail and high structural fidelity.

StyleSwin: More recently, Zhang et al. [3] developed *StyleSwin*, a novel approach that leverages the Swin Transformer architecture in conjunction with style modulation. Unlike its predecessors, StyleSwin does not employ explicit masking; instead, it capitalizes on local attention mechanisms (window-based self-attention) and style injection to effectively model both the global layout and local texture of high-resolution images. Despite the absence of masking, StyleSwin achieves superior generative performance across several datasets, reporting FID scores of 2.81 (FFHQ-256), 5.07 (FFHQ-1024), 3.25 (CelebA-HQ 256), 4.43 (CelebA-HQ 1024), and 2.95 (LSUN-Church 256). These results not only surpass the performance of StyleGAN2 in many cases but also demonstrate the scalability and efficacy of transformer-based GANs in high-resolution synthesis tasks.

Collectively, these studies highlight the versatility and effectiveness of transformer architectures in generative modeling. Each approach contributes unique methodological insights—ranging from latent space manipulation and hierarchical feature learning to style-aware attention mechanisms—that advance the current state of image synthesis. The convergence of transformers and GANs thus presents a promising trajectory for future research in controllable, high-resolution generative modeling.

TABLE I: Comparison of Transformer-based Models for Image Editing

Reference	Model	Classification	Dataset	Year	Masking	Region of Interest (ROI)	Result
Park J. et al. [1]	*Styleformer*	Transformer-based latent space editing	CIFAR-10, CelebA, STL-10	2022	Latent mask in transformer block	Style-control in latent space	Competitive performance in preserving style and structure 10.00 (IS), 3.92 (FID) 11.01 (IS), 15.17 (FID)
Zhao L. et al. [2]	*HIT (Hierarchical Transformer)*	Patch-wise hierarchical transformer	FFHQ, CelebA-HQ	2021	Hierarchical latent masking	Multi-level feature hierarchy	High-quality high-res generation 2.95 (FID), 3.39 (FID)
Zhang B. et al. [3]	*StyleSwin*	Swin Transformer with Style Modulation	FFHQ, CelebA-HQ, LSUN-Church	2021	NA	NA	Outperforms StyleGAN2 on image fidelity FID: **2.81** (FFHQ-256), **5.07** (FFHQ-1024) FID: **3.25** (256), **4.43** (1024) FID: **2.95** (256)

III. Methodology

Our Modified TransGAN model consists of approximately **87 million trainable parameters for the Generator** and a comparable scale for the Discriminator, emphasizing its expressive capacity for medical image generation.

A. Data Selection

We utilize the publicly available ISIC skin lesion dataset from the International Skin Imaging Collaboration (ISIC) archive. Specifically, we draw samples from ISIC 2019 and related collections, which contain a broad range of dermoscopic images representing common skin lesions, including melanoma, nevus, and seborrheic keratosis. These images are high-resolution, RGB, and annotated by dermatology experts, making the dataset a strong benchmark for research in skin cancer detection, classification, and segmentation.

For our generative task, we curated a filtered subset of the dataset that maintains consistency in resolution and image quality. All images were resized to 32×32 pixels, normalized to the range $[-1, 1]$, and converted into a format suitable for GAN training. We excluded images with severe artifacts, occlusions, or poor lighting to ensure clean inputs. This preprocessing ensures that our model learns meaningful lesion features and generates high-quality, diverse samples for medical imaging augmentation.

B. Model Architecture

We build upon the TransGAN architecture proposed by Yifan Jiang et al. [12], which introduced a convolution-free GAN design by leveraging transformer blocks. While the original model performed well on natural image datasets, it lacked spatially localized feature control and resolution-specific optimization—crucial for medical images like skin lesions. Therefore, we propose the Modified TransGAN, incorporating CNN components and positional embeddings to improve spatial precision and training stability.

1) Generator Architecture: Our Generator begins with a random latent input vector of size 512, which is passed through an MLP to produce an initial feature map of shape $(8 \times 8 \times 1024)$. Unlike the original TransGAN, where PixelShuffle layers are used for upsampling, we replace them with CNN Transpose (deconvolution) layers to enable better spatial feature propagation. Additionally:

- Positional Embedding is applied at the beginning of every transformer block to improve spatial consistency across resolutions.
- The final linear unflattening layer used in the original model is replaced with a convolutional layer for sharper image generation.

Figure 1 illustrates the model's training stability across epochs. The complete modified Generator architecture is summarized in Table II, while the original TransGAN design is shown in Table III for comparison.

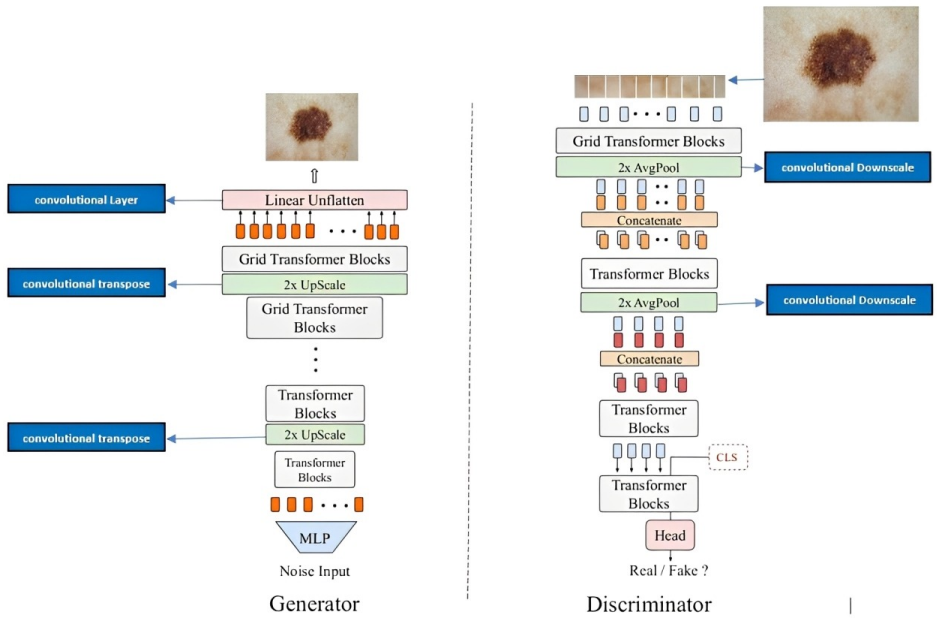


Fig. 1: Model Architecture of Modified TransGAN

TABLE II: Modified Generator Architecture

Layer	Input	Output
MLP + Reshape	$(B, 512)$	$(8 \times 8) \times 1024$
PosEmbedding + Block 1	$(8 \times 8) \times 1024$	$(8 \times 8) \times 1024$
Block 2	$(8 \times 8) \times 1024$	$(8 \times 8) \times 1024$
Block 3	$(8 \times 8) \times 1024$	$(8 \times 8) \times 1024$
Block 4	$(8 \times 8) \times 1024$	$(8 \times 8) \times 1024$
Block 5	$(8 \times 8) \times 1024$	$(8 \times 8) \times 1024$
CNN Upsample	$(8, 8, 1024)$	$(16, 16, 256)$
Flatten	$(16, 16, 256)$	$(16 \times 16) \times 256$
PosEmbed + Block 1	$(16 \times 16) \times 256$	$(16 \times 16) \times 256$
Block 2	$(16 \times 16) \times 256$	$(16 \times 16) \times 256$
Block 3	$(16 \times 16) \times 256$	$(16 \times 16) \times 256$
Block 4	$(16 \times 16) \times 256$	$(16 \times 16) \times 256$
CNN Upsample	$(16, 16, 256)$	$(32, 32, 64)$
Flatten	$(32, 32, 64)$	$(32 \times 32) \times 64$
PosEmbed + Block 1	$(32 \times 32) \times 64$	$(32 \times 32) \times 64$
Block 2	$(32 \times 32) \times 64$	$(32 \times 32) \times 64$
Final Conv Layer	$(32, 32, 64)$	$32 \times 32 \times 3$

TABLE III: Original TransGAN Generator

Layer	Input	Output
MLP	512	$(8 \times 8) \times 1024$
Block 1	$(8 \times 8) \times 1024$	$(8 \times 8) \times 1024$
Block 2	$(8 \times 8) \times 1024$	$(8 \times 8) \times 1024$
Block 3	$(8 \times 8) \times 1024$	$(8 \times 8) \times 1024$
Block 4	$(8 \times 8) \times 1024$	$(8 \times 8) \times 1024$
Block 5	$(8 \times 8) \times 1024$	$(8 \times 8) \times 1024$
PixelShuffle	$(8 \times 8) \times 1024$	$(16 \times 16) \times 256$
Block 6	$(16 \times 16) \times 256$	$(16 \times 16) \times 256$
Block 7	$(16 \times 16) \times 256$	$(16 \times 16) \times 256$
Block 8	$(16 \times 16) \times 256$	$(16 \times 16) \times 256$
Block 9	$(16 \times 16) \times 256$	$(16 \times 16) \times 256$
PixelShuffle	$(16 \times 16) \times 256$	$(32 \times 32) \times 64$
Block 10	$(32 \times 32) \times 64$	$(32 \times 32) \times 64$
Block 11	$(32 \times 32) \times 64$	$(32 \times 32) \times 64$
Linear Layer	$(32 \times 32) \times 64$	$32 \times 32 \times 3$

2) Discriminator Architecture: The original TransGAN Discriminator uses linear flattening, transformer blocks, and average pooling to extract features at multiple resolutions. It processes the input through a sequence of transformer blocks and finally appends a CLS token to classify the image.

Our Modified Discriminator improves on this by integrating convolutional operations and positional embeddings, while retaining the transformer-based decision-making core. We also adopt DiffAug for online data augmentation to regularize training. A detailed breakdown of both the modified and original discriminator architectures is presented in Table IV and Table V, respectively.

TABLE IV: Modified Discriminator Architecture

Layer	Input Shape	Output Shape
Data Augmentation (DiffAug)	$(32, 32, 3)$	$(32, 32, 3)$
Patch Embedding (Conv2D)	$(32, 32, 3)$	$(16, 16, 192)$
Reshape + PosEmbedding	$(16 \times 16, 192)$	$(256, 192)$
Transformer Blocks × 3	$(256, 192)$	$(256, 192)$
AvgPool + Reshape	$(16, 16, 192)$	$(8 \times 8, 192)$
Downsample (Conv2D)	$(32, 32, 3)$	$(8, 8, 192)$
Reshape	$(8 \times 8, 192)$	$(64, 192)$
Concatenate with x1	$(64, 192)$ + $(64, 192)$	$(64, 384)$
PosEmbedding + Blocks × 3	$(64, 384)$	$(64, 384)$
CLS Token + Transformer Block	$(65, 384)$	$(65, 384)$
LayerNorm + Dense	$(65, 384)$	$(1,)$ (logit)

TABLE V: Original Discriminator Architecture

Layer	Input Shape	Output Shape
Linear Flatten	$32 \times 32 \times 3$	$(16 \times 16) \times 192$
Block 1	$(16 \times 16) \times 192$	$(16 \times 16) \times 192$
Block 2	$(16 \times 16) \times 192$	$(16 \times 16) \times 192$
Block 3	$(16 \times 16) \times 192$	$(16 \times 16) \times 192$
AvgPooling	$(16 \times 16) \times 192$	$(8 \times 8) \times 192$
Concatenate	$(8 \times 8) \times 192$	$(8 \times 8) \times 384$
Block 4	$(8 \times 8) \times 384$	$(8 \times 8) \times 384$
Block 5	$(8 \times 8) \times 384$	$(8 \times 8) \times 384$
Block 6	$(8 \times 8) \times 384$	$(8 \times 8) \times 384$
Add CLS Token	$(8 \times 8) \times 384$	$(8 \times 8 + 1) \times 384$
Block 7	$(8 \times 8 + 1) \times 384$	$(8 \times 8 + 1) \times 384$
CLS Head	1×384	1

C. Hyperparameter Tuning

Since TransGAN was originally designed for general-purpose natural image synthesis, several domain-specific hyperparameter adaptations were necessary to optimize performance for medical image generation, specifically for skin lesion synthesis using the ISIC dataset.

The following adjustments were made to tailor the model to the unique texture, color distribution, and structure found in dermatological images:

- **Number of Transformer Layers:** To adequately capture the intricate details of skin lesion patterns, we experimented with a range of 6–12 transformer layers across different resolution stages. A higher number of layers enabled deeper feature modeling, while balancing computational overhead.
- **Attention Heads:** The number of attention heads per transformer block was tuned based on resolution and GPU memory constraints. Fewer heads were used in early blocks to reduce complexity while preserving performance. Our final configuration used [4, 4, 4] heads across the resolution stages.
- **MLP Dimensions (Generator and Discriminator):** Generator MLP: [512, 512, 512], Discriminator MLP: [768, 1024, 1024]. These sizes allowed sufficient representational capacity while maintaining training stability.
- **Learning Rate:** Set to 2×10^{-4} with scheduled decay to stabilize convergence and avoid overshooting in early training phases.
- **Optimizer:** We employed Adam optimizer with the following settings: $\beta_1 = 0.0$ (to allow faster adaptation to gradient updates), $\beta_2 = 0.99$ (to maintain stable second-moment estimates). Generator learning rate ($g\_learning\_rate$): 0.0001, fine-tuned for smoother image generation and to reduce adversarial noise.

- **Discriminator Configuration:** Feature dimensions: d_dim : $[192, 64]$, layered based on spatial resolution to balance granularity and memory efficiency. Depth: d_depth : $[2, 2]$, indicating two transformer blocks at each major resolution stage for the discriminator.
- **Batch Size:** A batch size of 64 was used, selected as the maximum feasible size given available GPU memory. This batch size ensures diversity during each training iteration without overloading hardware resources.

For our generative task, we curated a subset of the dataset with consistent resolution and quality to maintain uniformity in training. All images were resized to **32×32** pixels, normalized to the range $[-1, 1]$, and converted to a standard format suitable for GAN training. We excluded images with significant artifacts, occlusions, or poor illumination to ensure data quality. This preprocessing pipeline ensures that the model is trained on clean and representative samples of various lesion types.

D. Evaluation Metrics

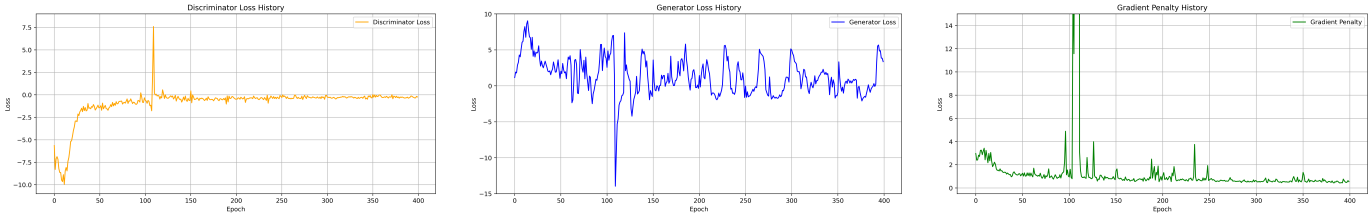
To assess the quality and diversity of generated images, we employed two commonly used GAN evaluation metrics:

- **Fréchet Inception Distance (FID)**: Measures the distance between feature distributions of real and generated images.

In addition to these metrics, **qualitative visual inspection** was conducted to assess visual fidelity and lesion realism. We also evaluated the **impact of synthetic data augmentation** on a downstream classification model trained to distinguish between lesion types.

IV. Results

The performance of the proposed model was assessed by analyzing generator loss, discriminator loss, and gradient penalty over the course of 400 training epochs. As shown in Table VI, the generator loss peaked at 9.213, dropped as low as −13.948, and had a mean value of 1.226. During the first 100 epochs, this loss exhibited significant fluctuation—common in GANs—but stabilized thereafter, consistently falling between −2 and +5, as illustrated in Figure 2b. Similarly, the discriminator loss reached a maximum of 7.596, a minimum of −9.943, and averaged −0.647. After 130 epochs, it became stable and oscillated in a narrow band between 0 and −0.5 (Figure 2a). The gradient penalty, which enforces the Lipschitz condition for WGAN stability, initially spiked (maximum 5562.82) but soon converged to a manageable average of 14.62, with a minimum of 0.383, confirming proper regularization (Figure 2c). These patterns indicate that both generator and discriminator reached a balanced and stable adversarial state. The final model not only achieved a Fréchet Inception Distance (FID) of 7.5, but also generated high-fidelity lesion images (Figure 3b) closely resembling real samples (Figure 3a), demonstrating its practical potential for dermatological image augmentation.



(a) Discriminator Loss (b) Generator Loss (c) Gradient Penalty

Fig. 2: Training curves showing the evolution of loss components during GAN training.

(a) Original Images (b) Generator Images

Fig. 3: Comparison between real and generated lesion images.

TABLE VI: Training loss statistics over 400 epochs for generator, discriminator, and gradient penalty.

Loss Type	Max	Min	Mean
Generator Loss	9.213	-13.948	1.226
Discriminator Loss	7.596	-9.943	-0.647
Gradient Penalty	5562.82	0.383	14.62

V. Discussion

This study presents a domain-specific adaptation of the transformer-based TransGAN architecture tailored for skin lesion image synthesis using the ISIC dataset. The original TransGAN model, while innovative in its convolution-free design and pure transformer backbone, was primarily developed for natural image synthesis and thus lacked several features critical for the medical imaging domain. Our modifications—replacing pixel shuffle upsampling with CNN transpose layers, introducing CNN-based average pooling, embedding positional encodings before every transformer block, and substituting the generator's final linear unflatten layer with a convolutional layer—significantly enhanced the model's ability to capture local texture and spatial coherence in high-detail skin lesion images.

The training behavior observed reflected typical challenges associated with GANs, particularly in medical imaging where subtle structural and textural nuances are essential. The fluctuations in the generator loss during the first 100 epochs followed by stabilization indicate a period of adversarial balance seeking. Similarly, the discriminator's convergence after about 130 epochs suggests a robust adversary learning to distinguish synthetic from real images effectively. The gradient penalty's role in stabilizing training and preventing mode collapse was evident, as the penalty initially peaked but maintained a controlled average thereafter, ensuring compliance with Lipschitz continuity constraints.

Achieving an FID score of 7.5 demonstrates the strong fidelity and diversity of the synthetic images relative to real data. This quantitative metric, coupled with qualitative visual inspection, confirms that the synthesized images retain the complexity and variability of real skin lesions, an important factor in building generalizable diagnostic models. This is particularly valuable given the data scarcity and imbalance often encountered in melanoma datasets, where high-quality synthetic images can help augment training sets, mitigate overfitting, and improve model robustness. Additionally, the architectural improvements permitted a reduction in the number of attention heads per transformer block without loss of performance, which is significant for computational efficiency. This reduction makes the model more feasible for deployment in research and clinical environments where GPU resources may be limited. Nonetheless, this study has some limitations. The image resolution was constrained to 32×32 pixels to balance computational resources and training stability, which may limit fine-grained feature representation important in clinical diagnostics. Future work should investigate scaling this approach to higher resolutions. Furthermore, while FID provides a useful proxy for image quality, clinical validation with dermatologists is necessary to assess the utility of synthetic images in real-world diagnostic workflows. Exploring advanced transformer variants such as hierarchical transformers or integrating multimodal data (e.g., clinical metadata alongside images) could further improve synthesis quality and clinical relevance. Moreover, extending this approach to other medical imaging modalities—such as histopathology, radiology, or ophthalmology—may reveal its broader applicability in medical image augmentation.

VI. Conclusion

In conclusion, we have successfully adapted the transformer-based TransGAN model for skin lesion image synthesis by incorporating CNN operations and positional embeddings tailored to the dermatological image domain. Our modified architecture achieved stable training dynamics and a competitive FID score of 7.5 on the ISIC melanoma dataset, generating high-fidelity synthetic images that closely resemble real samples. This research underscores the importance of domain-specific architectural modifications when applying advanced generative models to specialized fields such as medical imaging. The ability to generate realistic, diverse synthetic skin lesion images holds promise for mitigating dataset limitations, enhancing training data for diagnostic models, and ultimately improving automated melanoma detection. Future directions include scaling synthesis to higher resolutions, performing rigorous clinical validation, integrating multimodal information, and exploring transformer enhancements to further boost generative capabilities. The demonstrated success of this approach offers a promising pathway for leveraging transformer-based GANs in the broader landscape of medical image analysis and augmentation.

Future Work

Future research directions include integrating TransGAN-generated data into training pipelines for skin lesion classification tasks and measuring any improvements in model performance. Investigating hybrid architectures that combine convolutional and transformer-based layers could also enhance synthesis quality. Applying the model to other domains such as histopathology or radiology may help assess its generalizability. Moreover, exploring semi-supervised or conditional GAN variants may allow for controlled generation of specific lesion types such as melanoma or benign cases.

This work highlights the potential of transformer-based GANs like TransGAN in medical image synthesis and lays the foundation for their integration into real-world dermatological AI systems.

REFERENCES

[1] Min Park, Hyeongwoo Kim, Jaesik Lee, Minsu Cho, and In So Kweon, "Styleformer: Transformer-based GAN for controlled image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12972–12982.

[2] Shangzhe Zhao, Yiming Liu, Chaoyang Pang, and Stephen Lin, "HiT: Hierarchical transformers for high-resolution image synthesis," in *Advances in Neural Information Processing Systems*, 2021.

[3] Jianmin Zhang et al., "StyleSwin: Transformer-based GAN for High-Resolution Image Synthesis," *arXiv preprint arXiv:2112.10762*, 2021.

[4] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

[5] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C. Courville. *Improved training of Wasserstein GANs.* In Advances in Neural Information Processing Systems, pages 5767–5777, 2017.

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. *Large scale GAN training for high fidelity natural image synthesis.* arXiv preprint arXiv:1809.11096, 2018.

[7] Karol Kurach, Mario Lucic, Xiaohua Zhai, Marcin Michalski, and Sylvain Gelly. *A large-scale study on regularization and normalization in GANs.* In International Conference on Machine Learning, pages 3581–3590. PMLR, 2019.

[8] Kevin Roth, Aurelien Lucchi, Sebastian Nowozin, and Thomas Hofmann. *Stabilizing training of generative adversarial networks through regularization.* In Advances in Neural Information Processing Systems, pages 2018–2028, 2017.

[9] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. *Least squares generative adversarial networks.* In Proceedings of the IEEE International Conference on Computer Vision, pages 2794–2802, 2017.

[10] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. *Improved techniques for training GANs.* arXiv preprint arXiv:1606.03498, 2016.

[11] Mario Lucic, Karol Kurach, Marcin Michalski, Olivier Bousquet, and Sylvain Gelly. *Are GANs created equal? A large-scale study.* In Proceedings of the 32nd International Conference on Neural Information Processing Systems, pages 698–707, 2018.

[12] Y. Jiang, S. Chang, and Z. Wang. *TransGAN: Two pure transformers can make one strong GAN, and that can scale up.* arXiv.org, https://arxiv.org/abs/2102.07074 (accessed Jun. 26, 2025).

[13] *ISIC Archive.* Isic-archive.com, 2016. https://api.isic-archive.com/collections/74/ (accessed Jun. 26, 2025).