Medical Imaging Complexity and its Effects on GAN Performance

Anonymous Author(s)
Affiliation
Address
email

Abstract

The proliferation of machine learning models in diverse clinical applications has led to a growing need for high-fidelity, medical image training data. Such data is often scarce due to cost constraints and privacy concerns. Alleviating this burden, medical image synthesis via generative adversarial networks (GANs) emerged as a powerful method for synthetically generating photo-realistic images based on existing sets of real medical images. However, the exact image set size required to efficiently train such a GAN is unclear. In this work, we experimentally establish benchmarks that measure the relationship between a sample dataset size and the fidelity of the generated images, given the dataset's distribution of image complexities. We analyze statistical metrics based on delentropy, an image complexity measure rooted in Shannon's entropy in information theory. For our pipeline, we conduct experiments with two state-of-the-art GANs, StyleGAN 3 and SPADE-GAN, trained on multiple medical imaging datasets with variable sample sizes. Across both GANs, general performance improved with increasing training set size but suffered with increasing complexity.

1 Introduction

Machine learning in healthcare is a rapidly growing field with countless applications [28] including disease diagnosis [24], clinical treatment [26], drug development [19], and mental health [7]. The machine learning models driving these advances require the collection of high-quality, annotated medical training data, which persists as an arduous task due to privacy concerns surrounding sensitive patient data [23] and the time-intensive nature of labeling [5]. To address these issues, synthetic data—artificially generated information mimicking real-world data—has surfaced as a promising solution [8].

Currently, generative adversarial networks (GANs) remain one of the leading approaches to synthetic data generation [17]. Since its inception in 2014 [6], GANs have gained increasing attention in the medical research community due to their ability to synthesize medical images [29]. However, achieving results with high fidelity remains a difficult task factoring the lack of medical data and prevalence of smaller datasets in the medical domain. With limited data, a GAN's efficacy is directly affected with consequences including mode collapse, where the generator produces a limited variety of outputs [20], and overfitting, where the GAN replicates training data rather than generalizing from it [33].

Various papers such as Wang et al. [32]'s transfer learning and Robb et al. [25]'s Few-Shot GAN (FSGAN) have addressed these issues as architecture-centric approaches, achieving increased training efficiency only as a result of the changes in the GAN's structure. However, such approaches are ineffective when making alterations to a GAN's internal structure are not feasible and when time constraints are present. As such, a data-centric approach by providing the GAN with the optimal

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

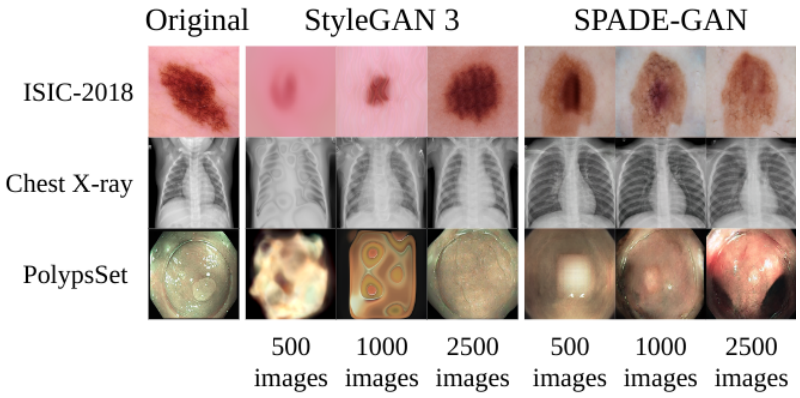

Figure 1: Comparison between original images and synthetic images from StyleGAN 3 and SPADE-GAN based on variable image set sizes.

amount of data to produce high-quality results is more appropriate. Nevertheless, the exact sample set size required to train state-of-the-art GANs is obscure.

In this study, we introduce a data-centric optimization method to create efficient GAN training for medical image synthesis. Our approach investigates how the image complexity distribution of a medical image dataset can be utilized as a measure of training difficulty for a GAN. By doing so, we can ascertain a correlation between the image complexities of the training images and the optimal training set sizes by establishing benchmarks that evaluate the relationship between a sample training set size and the fidelity of the generated images. We hypothesize that given a dataset of a specific image complexity distribution, healthcare professionals can reference the closest image fidelity curve to identify the optimal amount of experimental trials to produce superlative results. Ultimately, our approach can avoid both undertraining and wasteful overtraining by constructing a data-efficient, GAN training pipeline.

2 Background

Generative Adversarial Networks (GANs) Introduced by Goodfellow et al. [6], GANs are a class of generative models that consist of two convolutional neural networks: a generator G, which aims to transform its latent variable distribution $p(z)$ to closely resemble the training data distribution $p(x)$, and a discriminator D, which differentiates between the ground truth and data generated by G. Training is an adversarial process where G attempts to deceive D into classifying its outputs as real. This two-player minimax game is represented by the following loss function:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\log D(\mathbf{x})\right] \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}\left[\log(1 - D(G(\mathbf{z})))\right]. \tag{1}$$

Many papers have tried to address data scarcity and computational costs in GAN training architecturally. One approach proposed by Wang et al. [32] is transfer learning, which consists of fine-tuning a pre-trained generator and discriminator to the desired domain. However, if the pre-trained models do not align well with the target domain, this could result in even higher data and computational demands [34]. Robb et al. [25] proposed another solution through their Few-Shot GAN (FSGAN), achieving impressive adaptation even with extremely few training examples, albeit at the cost of prolonged training times. This results in the reduced quality and diversity of the synthetic data when time constraints are present.

Image Complexity Objectively, image complexity can be defined as the variety of features and details within an image. It has been shown that information entropy is a traditional, heuristic-based method of calculating the complexities of images in small-scale datasets [16].

Traditional information entropy, or Shannon entropy, is a foundational abstraction in information theory introduced by Shannon [27]. Used as a measure of uncertainty or "surprise" in data, it is the

variation in the distribution of pixel intensities of an image in grayscale format. The equation is defined as

$$H = -\sum_{i=0}^{n-1} p_i \log_b p_i, \tag{2}$$

where n denotes the number of gray levels (256 for 8-bit images), b stands for the logarithmic base (returning bits when $b = 2$), and p_i is the probability of a pixel having gray level i. However, although Shannon entropy considers compositional image information, it fails to account for spatial information, specifically the relationship between neighbouring pixels [4].

Another entropy-based metric, the Gray Level Co-Occurrence Matrix (GLCM), unlike Shannon's Entropy, is a measure of how often pairs of pixel values occur in a grayscale image distribution [9]. Taking into account local spatial information, the GLCM is useful for textural analysis tasks such as feature extraction for medical image segmentation [14]. The GLCM entropy can be represented as

$$H_g = -\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} p_{(i,j)} \log_b p_{(i,j)}, \tag{3}$$

where $p_{(i,j)}$ is the probability of a two pixels having gray levels i and j at a certain angle θ and distance d away from each other.

3 Methodology

3.1 Image Complexity Metric

We utilize Larkin's delentropy, a metric identical to the Shannon entropy and the GLCM, but incorporating a new density function known as the *deledensity* [15]. By analyzing the relationship between the local and global features of an image, delentropy accounts for an image's gradient vector field and pixel co-occurrence, encapsulating its spatial information as a whole. The deledensity, as a joint probability function is formulated as

$$p_{(i,j)} = \frac{1}{4WH} \sum_{w=0}^{W-1}\sum_{h=0}^{H-1} \delta_{i,d_x(w,h)}\delta_{j,d_y(w,h)}, \tag{4}$$

where d_x and d_y denote the derivative kernels in the x and y direction, δ is the Kronecker delta to describe the binning operation required to generate a histogram, and H and W is the image's dimensions (height and width) [13]. By obtaining this, we can then calculate delentropy as

$$DE = -\frac{1}{2}\sum_{i=0}^{I-1}\sum_{j=0}^{J-1} p_{(i,j)} \log_b p_{(i,j)}, \tag{5}$$

such that I and J represent the number of bins (discrete cells) in the 2D distribution, and the $\frac{1}{2}$ is derived from Papoulis' generalized sampling expansion [21].

To interpret this measure, yielding a high delentropy suggests an image has a high range of variation in pixel intensities and more sophisticated details. A low delentropy can be interpreted as a result of having a uniform distribution of pixel intensities, indicating simple structure and a less-detailed image.

Prior to any calculations, each image is preprocessed into an 8-bit, grayscale image. This ensures delentropy can be calculated in a consistent, single-channel format throughout each dataset.

3.2 GAN Selection

Core to the experimental approach was the selection of two state-of-the-art GANs, SPADE-GAN [22] and StyleGAN 3 [11] on which to run the experimental pipeline. These networks have been widely adopted by the medical image synthesis community and empirically observed to produce superior-quality medical images when compared to predecessor GANs [30]. StyleGAN 3's large community support and wide availability of its code repository along with its numerous configurations for different training settings was taken into account as well.

3.3 GAN Pipeline

Throughout the experiment, as our approach was data-centric, StyleGAN 3 and SPADE-GAN were run with the official, publicly available implementations with default hyperparameters and no augmentations to each network's architecture.

Preprocessing We first set all images to a consistent 512x512 resolution. As such, training parameters were based on the size of the preprocessed images, as documented in the official implementations. SPADE-GAN additionally relies on segmentation masks to produce synthetic data. We used pre-existing annotations for ISIC-2018 and the Polyps Set. Because the Chest X-ray dataset did not have such annotations, masks were generated using TorchXRayVision [3]. All experiments were performed on one NVIDIA A100 and three NVIDIA A40 GPUs.

Training and Generation The experimental pipeline is designed to identify the role of image dataset size in the image generation fidelity of selected GANs. To that end, for each GAN training run, all parameters are held constant with the exception of the image set size, which were set to 500, 1000, and 2500 images, randomly sampled from the same dataset for each experimental run, respectively. For StyleGAN 3, all experimental runs were trained for 100 epochs; for SPADE-GAN, 50 epochs. The trained adversarial network is then used to generate synthetic images, the fidelity of which is then evaluated for each training set size.

Evaluation The Fréchet Inception Distance (FID) [10] is a common metric used to evaluate the fidelity of the synthetically generated images for GANs [1]. Defined as the distance between the distributions of ground truth and the generated images respectively, in our paper we use the FID to assess the performance of the GAN (i.e. image fidelity) for each experimental run across both GANs. **A lower FID score signifies that a GAN is more proficient at generating synthetic data close to its target distribution**. From these data, we obtain fidelity curves for each dataset that describe how FID scores trend with increasing training set size.

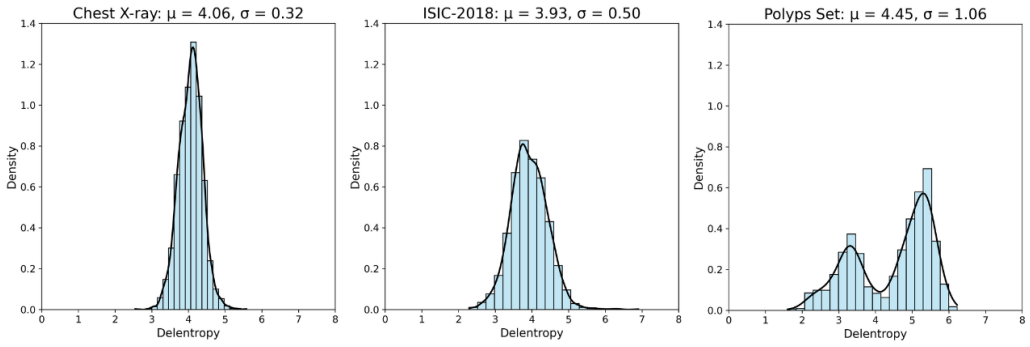

Figure 2: Delentropy distributions across each medical image dataset. A higher mean delentropy μ indicates a dataset with more complex images.

4 Experimental Results

Datasets We employ three medical imaging datasets: International Skin Imaging Collaboration 2018 Challenge (ISIC-2018) [2], Chest X-Ray Images (Chest X-ray) [12], and Colonoscopy Polyp

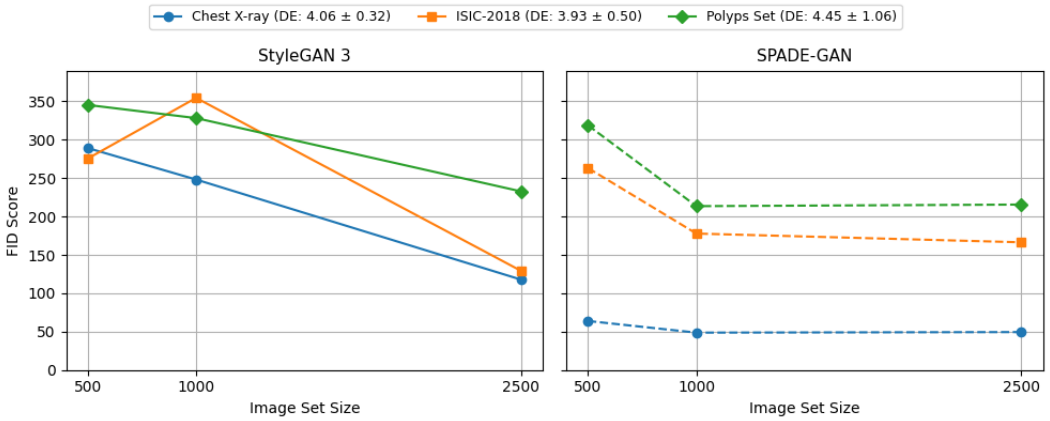

Figure 3: Fréchet Inception Distance (FID) curves comparing StyleGAN 3 and SPADE-GAN across each medical image dataset with varying sample sizes. Lower FID scores correspond to higher fidelity synthetic images.

Detection and Classification (Polyps Set) [31]. These datasets were chosen for their diversity in both perceptual complexity, ranging from relatively skin lesions to complex colon polyps, and imaging modality (dermoscopy vs. x-ray vs. colonoscopy).

We carry out delentropy calculations as described in Section 3 by using a publicly available implementation from Marchesoni [18]. To effectively capture the overall complexity of each image dataset, we capture each dataset's delentropy distribution as displayed in Fig. 2.

Across the experimental runs, FID scores consistently decreased with increasing dataset size. On StyleGAN 3, synthesized images that had been generated by a GAN trained on 2500 images exhibited an average FID score reduction of 48% when compared to those generated by a StyleGAN 3 that had been trained on a mere 500 images (Fig. 3). SPADE-GAN experienced an analogous 31% FID score reduction on average, though it is worth noting that FID reduction plateaued after only 1000 training images.

Comparing both Fig. 2 and Fig. 3, one can see a general relationship between the delentropy distribution and the training performance of both GANs. As the spread of image complexities increases from a slender, peaked distribution to a broader, bimodal one, we see a corresponding increase in FID scores for each dataset sample size. The Chest X-ray dataset with the most homogeneous image complexities shown by a tall and narrow distribution, yields the lowest FID score after being trained for 2500 images, indicating that both GANs had easier training runs with this dataset. On the contrast, the Polyps Set—the dataset with the widest distribution and multiple complexity peaks—correlates with the highest FID scores for each dataset sample size, which suggests that the GAN was faced with a more challenging and unstable training run. Ultimately, this pattern shows a general inverse relationship—GAN performance decreases with an increasing spread of image complexities within a dataset.

5 Discussion

SPADE-GAN outperformed StyleGAN 3 across *all datasets and training sizes*, with FID scores averaging 33% lower, likely due to its architecture that incorporates segmentation masks for structural information, whereas StyleGAN 3 trained on raw image data alone, making it more difficult to generalize to high-delentropy datasets. Furthermore, ISIC-2018 being an outlier can be attributed to its fluctuations in image complexity, reflected by the standard deviation in delentropy (Fig. 2). Despite having a lower mean delentropy, its spread likely resulted in difficulties in GAN training and learning the images' distribution, contrasting with the Chest X-ray dataset.

While the experimental results generally reflected an intuitive understanding of how image complexity and training data influence GAN training, the FID curves provide insightful details, offering a deeper perspective on these effects. SPADE-GAN exhibits both better quality results than StyleGAN 3 in the

form of lower FID scores and more consistent training as evidenced by the smooth, non-overlapping FID curves (Fig. 3). As aforementioned, performance plateaued after 1000 training images, suggesting that additional training data past that point may not help increase GAN performance as measured by FID score. This is also apparent in the generated images themselves, which exhibit little perceptual difference between those generated after 1000 training images and those generated after 2500 (Fig. 1). Contrast this with the StyleGAN 3 curves, which do not reach any noticeable plateau between 500 and 2500 training images. In fact, the increasingly negative slope values of the StyleGAN 3 graphs imply that StyleGAN 3 begins to better capture the images' features at a point past 1000 images, the exact whereabouts of which would need to be determined by a separate study.

The FID curves generated by this set of experiments set up a useful benchmark to which other potential training image data sets can be compared. For training sets that are of similar delentropy distributions and used to train StyleGAN 3 or SPADE-GAN, it is not unreasonable to predict that their training curves will be similar to those represented in Fig. 3, though many more training set sizes and image sets are required before a truly comprehensive representation can be reached.

Broader Impacts Our research on GANs for medical image synthesis may have positive and negative societal implications. On the positive side, it can enhance healthcare outcomes by improving the training of machine learning models with realistic synthetic data, therefore protecting patient privacy. Contrarily, potential negative impacts include the risk of malicious use for generating fraudulent synthetic data and possibility of reinforcing biases due to a lack of diversity of representing patient populations. These considerations demonstrate the importance of addressing both the benefits and potential risks associated with the use of GANs in the medical domain.

6 Conclusion

In this work, we highlight the impact of image complexity on GAN performance in medical image synthesis. We empirically demonstrate the general inverse relationship of how higher image complexity leads to poorer image fidelity results and performance in GANs. By displaying FID curves, we show healthcare professionals the possibility for the use of our benchmarks to gauge an estimate of data training requirements to achieve desirable results based on image complexity.

7 Limitations

Due to limited resources, experiments were only run on 500, 1000, and 2500 training images, leading to coarse-grained results. An extended study with a larger range and finer-grained increments would better elucidate exactly how FID scores respond to changes in training image dataset size. The use of FID scores as an evaluation metric also has its limitations, not necessarily correlating with human perceptual interpretations, something that is extremely important in the medical field where human doctors are still largely the source of truth. Skandarani et al. [29] show that a lower FID score is not a good measure of how well synthetic images can perform on a downstream task. More research involving multiple evaluations of the same experimental setup is required.

References

[1] Ali Borji. Pros and cons of gan evaluation measures: New developments. *Computer Vision and Image Understanding*, 215:103329, 2022. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2021.103329. URL `https://www.sciencedirect.com/science/article/pii/S1077314221001685`.

[2] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), 2018. URL `https://arxiv.org/abs/1902.03368`. Licensed under the Creative Commons Attribution-NonCommercial (CC-BY-NC) license.

[3] Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. TorchXRayVision: A library of chest X-ray datasets and models. In *Medical Imaging with Deep Learning*, 2022. URL `https://github.com/mlmed/torchxrayvision`.

[4] Peichao Gao, Zhilin Li, and Hong Zhang. Thermodynamics-based evaluation of various improved shannon entropies for configurational information of gray-level images. *Entropy*, 20(1), 2018. ISSN 1099-4300. doi: 10.3390/e20010019. URL https://www.mdpi.com/1099-4300/20/1/19.

[5] Andrew Gilbert, Maciej Marciniak, Cristobal Rodero, Pablo Lamata, Eigil Samset, and Kristin Mcleod. Generating synthetic labeled data from existing anatomical models: An example with echocardiography segmentation. *IEEE Transactions on Medical Imaging*, 40(10):2783–2794, 2021. doi: 10.1109/TMI.2021.3051806.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[7] Sarah Graham, Colin Depp, Ellen E Lee, Camille Nebeker, Xin Tu, Ho-Cheol Kim, and Dilip V Jeste. Artificial intelligence for mental health and mental illnesses: an overview. *Current psychiatry reports*, 21: 1–18, 2019.

[8] Xu Guo and Yiqiang Chen. Generative ai for synthetic data generation: Methods, challenges and the future. *arXiv preprint arXiv:2403.04190*, 2024.

[9] Robert M. Haralick, K. Shanmugam, and Its'Hak Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621, 1973. doi: 10.1109/TSMC.1973.4309314.

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. URL https://arxiv.org/abs/1706.08500.

[11] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan, 2020. URL https://arxiv.org/abs/1912.04958. Licensed under the Nvidia Source Code License.

[12] Daniel Kermany, Kang Zhang, and Michael Goldbaum. Labeled optical coherence tomography (oct) and chest x-ray images for classification, 2018. URL https://doi.org/10.17632/rscbjbr9sj.2. Licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

[13] T. M. Khan, S. S. Naqvi, and E. Meijering. Leveraging image complexity in macro-level neural network design for medical image segmentation. *Scientific Reports*, 12, 2022. doi: 10.1038/s41598-022-26482-7. URL https://doi.org/10.1038/s41598-022-26482-7.

[14] Z. Faizal khan and Sultan Refa Alotaibi. Computerised segmentation of medical images using neural networks and glcm. In *2019 International Conference on Advances in the Emerging Computing Technologies (AECT)*, pages 1–5, 2020. doi: 10.1109/AECT47998.2020.9194196.

[15] Kieran G. Larkin. Reflections on shannon information: In search of a natural information-entropy for images, 2016. URL https://arxiv.org/abs/1609.01117.

[16] Shipeng Liu, Liang Zhao, Dengfeng Chen, and Zhanping Song. Contrastive learning for image complexity representation, 2024. URL https://arxiv.org/abs/2408.03230.

[17] Yingzhou Lu, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, Tianfan Fu, and Wenqi Wei. Machine learning for synthetic data generation: A review, 2024. URL https://arxiv.org/abs/2302.04062.

[18] Franco Marchesoni. How do i know if an image after image enhancement is better than before?, March 2023. URL https://ai.stackexchange.com/questions/39483/how-do-i-know-if-image-after-image-enhancement-is-better-than-before-image-pre.

[19] Galia Nordon, Gideon Koren, Varda Shalev, Eric Horvitz, and Kira Radinsky. Separating wheat from chaff: joining biomedical knowledge and patient data for repurposing medications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9565–9572, 2019.

[20] Zhaoqing Pan, Weijie Yu, Xiaokai Yi, Asifullah Khan, Feng Yuan, and Yuhui Zheng. Recent progress on generative adversarial networks (gans): A survey. *IEEE access*, 7:36322–36333, 2019.

[21] A. Papoulis. Generalized sampling expansion. *IEEE Transactions on Circuits and Systems*, 24(11): 652–654, 1977. doi: 10.1109/TCS.1977.1084284.

[22] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization, 2019. URL https://arxiv.org/abs/1903.07291. Licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license.

[23] Vasileios C. Pezoulas, Dimitrios I. Zaridis, Eugenia Mylona, Christos Androutsos, Kosmas Apostolidis, Nikolaos S. Tachos, and Dimitrios I. Fotiadis. Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and Structural Biotechnology Journal*, 23: 2892–2910, 2024. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2024.07.005. URL https://www.sciencedirect.com/science/article/pii/S2001037024002393.

[24] SM Rahman, Sifat Ibtisum, Ehsan Bazgir, and Tumpa Barai. The significance of machine learning in clinical disease diagnosis: A review. *arXiv preprint arXiv:2310.16978*, 2023.

[25] Esther Robb, Wen-Sheng Chu, Abhishek Kumar, and Jia-Bin Huang. Few-shot adaptation of generative adversarial networks. *arXiv preprint arXiv:2010.11943*, 2020.

[26] Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. Pre-training of graph augmented transformers for medication recommendation. *arXiv preprint arXiv:1906.00346*, 2019.

[27] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3): 379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

[28] Zheyuan Ryan Shi, Claire Wang, and Fei Fang. Artificial intelligence for social good: A survey. *arXiv preprint arXiv:2001.01818*, 2020.

[29] Youssef Skandarani, Pierre-Marc Jodoin, and Alain Lalande. GANs for Medical Image Synthesis: An Empirical Study. *Journal of Imaging*, 9(3):69, 2023. ISSN 2313-433X. doi: 10.3390/jimaging9030069. URL https://www.mdpi.com/2313-433X/9/3/69.

[30] Youssef Skandarani, Pierre-Marc Jodoin, and Alain Lalande. Gans for medical image synthesis: An empirical study. *Journal of Imaging*, 9, 2023. doi: 10.3390/jimaging9030069. URL https://www.mdpi.com/2313-433X/9/3/69.

[31] Guanghui Wang. Replication Data for: Colonoscopy Polyp Detection and Classification: Dataset Creation and Comparative Evaluations, 2021. URL https://doi.org/10.7910/DVN/FCBUOR. Licensed under the CC0 1.0 Universal (CC0 1.0) Public Domain Dedication.

[32] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost Van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: generating images from limited data. In *Proceedings of the European conference on computer vision (ECCV)*, pages 218–234, 2018.

[33] Ryan Webster, Julien Rabin, Loic Simon, and Frédéric Jurie. Detecting overfitting of deep generative networks via latent recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11273–11282, 2019.

[34] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

A Raw FID Scores

Dataset	Image Set Size	StyleGAN 3	SPADE-GAN
Chest X-ray	500	289.20	63.90
	1000	248.20	48.74
	2500	117.85	49.49
ISIC-2018	500	275.67	263.09
	1000	354.61	177.82
	2500	129.30	166.45
Polyps Set	500	345.42	318.76
	1000	328.27	213.54
	2500	232.78	215.61

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: They state our aim of establishing benchmarks for the relationship between dataset size and image fidelity in GAN-generated medical images, considering image complexity distribution in a variety of medical image datasets.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: The paper includes a dedicated "Limitations" section (Section 7) that discusses the constraints of the study. For instance, our limited range of training image set sizes and the limitations of using FID scores as the primary evaluation metric.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper is primarily based on empirical results and does not present theoretical results requiring proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides a detailed, step-by-step process of the GAN pipeline including the information on the datasets used, preprocessing steps, the number of training image sample size, training iterations, GAN architectures, and evaluation metrics.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The paper provides open access to both the data and code in the supplementary material. Sufficient instructions are included to allow correct reproduction of the main experimental results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the datasets used, preprocessing steps, GAN architectures (StyleGAN 3 and SPADE-GAN), number of training images (500, 1000, 2500), and number of epochs for each GAN, outlined in the methodology (Section. 3).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: This paper presents FID scores for different experimental conditions, and it also reports the variability in the image complexities of each medical imaging dataset as shown in Fig. 2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The paper says that experiments were performed on "one NVIDIA A100 and three NVIDIA A40 GPUs.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The paper conforms with the NeurIPS Code of Ethics such as providing our publicly used datasets.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: This paper explains both the positive and negative broader effects under the discussion (Section 5).

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper does not release new models or datasets and poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: This paper properly credits used GAN models and datasets and can be found as a note in the reference of each asset.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not introduce new assets such as datasets or models.

    Guidelines:

    - The answer NA means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve any crowdsourcing or research with human subjects

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve any research with human subjects that would require IRB approval.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.