Feature Extraction for Generative Medical Imaging Evaluation: New Evidence Against an Evolving Trend

McKell Woodland[1,2][0000−0003−0995−3695], Austin Castelo[1][0009−0009−3015−3570], Mais Al Taie[1][0009−0000−9298−6644], Jessica Albuquerque Marques Silva[1][0000−0003−0616−982X], Mohamed Eltaher[1][0009−0008−2078−4718], Frank Mohn[1][0009−0001−9264−9725], Alexander Shieh[1][0009−0005−7450−2702], Suprateek Kundu[1][0000−0002−1767−4875], Joshua P. Yung[1][0000−0003−3752−5997], Ankit B. Patel[2,3][0000−0001−9678−496X], and Kristy K. Brock[1][0000−0001−9364−5040]

[1] The University of Texas MD Anderson Cancer Center, Houston TX 77030, USA
{mewoodland,ahcastelo,mmal1,jalbuquerque,meltaher,adshieh,skundu2, jyung,kkbrock}@mdanderson.org
[2] Rice University, Houston TX 77005, USA
ankit.patel@rice.edu
[3] Baylor College of Medicine, Houston TX 77030, USA

Abstract. Fréchet Inception Distance (FID) is a widely used metric for assessing synthetic image quality. It relies on an ImageNet-based feature extractor, making its applicability to medical imaging unclear. A recent trend is to adapt FID to medical imaging through feature extractors trained on medical images. Our study challenges this practice by demonstrating that ImageNet-based extractors are more consistent and aligned with human judgment than their RadImageNet counterparts. We evaluated sixteen StyleGAN2 networks across four medical imaging modalities and four data augmentation techniques with Fréchet distances (FDs) computed using eleven ImageNet or RadImageNet-trained feature extractors. Comparison with human judgment via visual Turing tests revealed that ImageNet-based extractors produced rankings consistent with human judgment, with the FD derived from the ImageNet-trained SwAV extractor significantly correlating with expert evaluations. In contrast, RadImageNet-based rankings were volatile and inconsistent with human judgment. Our findings challenge prevailing assumptions, providing novel evidence that medical image-trained feature extractors do not inherently improve FDs and can even compromise their reliability. Our code is available at https://github.com/mckellwoodland/fid-med-eval.

Keywords: Generative Medical Imaging · Fréchet Inception Distance.

1 Introduction

Fréchet Inception Distance (FID) is a commonly used metric for evaluating synthetic image quality [1]. It quantifies the Fréchet distance (FD) between two

Gaussian distribution curves fitted to embeddings of real and generated images. These embeddings are typically extracted from the penultimate layer of an InceptionV3 network trained on ImageNet. FID's utility has been demonstrated through its correlation with human judgment [2], sensitivity to distortions [1], capability to detect overfitting [3], and relative sample efficiency [3]. Nonetheless, the metric has faced criticism, including that the InceptionV3 network may only embed information relevant to ImageNet class discrimination [4,5].

Three approaches exist for adapting FID to medical imaging. The first involves using an InceptionV3 extractor trained on a large, publicly available medical dataset, such as RadImageNet, a database containing 1.35 million annotated computed tomography (CT), magnetic resonance imaging (MRI), and ultrasonography exams [6,7]. While a RadImageNet-based FD considers medically relevant features, its efficacy remains largely unexplored. One potential bias is that networks trained for disease detection may focus too heavily on small, localized regions [8] to evaluate an entire image's quality effectively. Additionally, RadImageNet-based FDs may not generalize to new medical modalities [7] or patient populations. Our novel comparison of RadImageNet-based FDs to human judgment revealed discrepancies, even on in-domain abdominal CT data.

The second approach utilizes self-supervised networks for feature extraction [9]. These networks are encouraging as they create transferable and robust representations [10], including on medical images [4]. Despite their promise, the lack of publicly available, self-supervised models trained on extensive medical imaging datasets has hindered their application. Our study is the first to employ self-supervised extractors for synthetic medical image evaluation. We find a significant correlation between an FD derived from an ImageNet-trained SwAV network (FSD) and medical experts' appraisal of image realism, highlighting the potential of self-supervision for advancing generative medical imaging evaluation.

The third approach employs a feature extractor trained on the dataset used to train the generative imaging model [11,12,13]. While advantageous for domain coherence, the algorithm designer creates the metric used to evaluate their algorithm, potentially resulting in unquantified bias. Moreover, the private and varied nature of these feature extractors poses challenges for reproducibility and benchmarking. Given these limitations, our study focuses on publicly available feature extractors.

Our study offers a novel comparison of generative model rankings created by ImageNet- and RadImageNet-trained feature extractors with expert judgment. Our main contributions are:

1. Demonstrating that ImageNet-based feature extractors consistently produce more realistic model rankings than their RadImageNet-based counterparts. This finding raises concerns about the prevalent practice of using medical image-trained feature extractors for generative model ranking without evaluating the efficacy of the proposed metric.
2. Identifying a significant correlation between an FD calculated with an ImageNet-trained SwAV network and expert assessments of image realism, demonstrating that FSD is a viable alternative to FID on medical images.

3. Benchmarking multiple data augmentation techniques designed to enhance generative performance within limited data domains on medical imaging datasets.
4. Introducing a novel method for evaluating visual Turing Tests (VTTs) via hypothesis testing, providing an unbiased measure of participant perception of synthetic image realism.

2 Methods

2.1 Generative Modeling

Four medical imaging datasets were used for generative modeling: the Segmentation of the Liver Competition 2007 (SLIVER07) dataset with 20 liver CT studies [14][1], the ChestX-ray14 dataset with 112,100 chest X-rays [15][2], the brain tumor dataset from the Medical Segmentation Decathlon (MSD) with 750 brain MRI studies [16,17][3], and the Automated Cardiac Diagnosis Challenge (ACDC) dataset with 150 cardiac cine-MRIs [18][4]. Multi-dimensional images were converted to two dimensions by extracting axial slices and excluding the slices with less than 15% nonzero pixels.

Four StyleGAN2 [19] models were trained per dataset, using either adaptive discriminator augmentation (ADA) [20], differentiable augmentation (DiffAugment) [21], adaptive pseudo augmentation (APA) [22], or no augmentation, to enable a comparison of synthetic quality. StyleGAN2 was chosen for its ability to produce high-fidelity medical images [2] and its readily available data augmentation implementations. While all of the data augmentation techniques were created to improve the performance of generative models on limited data domains, such as medical imaging, we are the first to benchmark the techniques on medical images. Each model was evaluated using the weights obtained at the end of 25,000 kimg (a kimg represents a thousand real images being shown to the discriminator), except for the MSD experiments, which were limited to 5,000 kimg due to training instability. Our code and trained model weights are available at https://github.com/mckellwoodland/fid-med-eval.

2.2 Human Evaluation

Human perception of model quality was assessed with one VTT per model. Each test comprised 20 randomly selected images with an equal number of real and generated images. Participants were asked to identify whether each image was real or generated and rate its realism on a Likert scale from 1 to 3 (1: "Not at all realistic," 2: "Somewhat realistic," and 3: "Very realistic"). The tests were administered to five specialists with medical degrees. In addition to the VTTs,

[1] https://sliver07.grand-challenge.org/
[2] https://nihcc.app.box.com/v/ChestXray-NIHCC
[3] http://medicaldecathlon.com/, CC-BY-SA 4.0 license.
[4] https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.html

three radiologists were shown 35 synthetic radiographs per ChestX-ray14 model and were asked to rank and provide a qualitative assessment of the models.

False positive rate (FPR) and false negative rate (FNR) were used to evaluate the VTTs. The FPRs represent the proportion of generated images that participants considered to be real. FPRs near 50% indicate random guessing. One-sided paired t tests were performed on the FPRs with α=.05 to benchmark the data augmentation techniques. For each VTT, the average Likert ratings of real and generated images were computed per participant. The difference between these average ratings (Diff) was then computed to compare the perceived realism of real and generated images. Two-sample Kolmogorov-Smirnov (KS) tests were conducted on the Likert ratings of the real and generated images with significance level α=.10 to determine whether the ratings came from the same distribution, indicating that the participants viewed the realism of the generated images to be equivalent to that of the real images. We are the first to use the difference in average Likert ratings and the KS test for generative modeling evaluation.

When taking a VTT, participants may be more likely to select either "real" or "generated" when uncertain. This bias causes the average FPR to not fully encapsulate whether participants can differentiate between real and generated images. We propose a novel method for evaluating VTTs via hypothesis testing to address this challenge. The method aims to demonstrate that the likelihood of a participant selecting "real" is the same for both real and generated images. We define the null hypothesis $\mathbb{P}(p \text{ guesses real} \mid \text{G}) = \mathbb{P}(p \text{ guesses real} \mid \text{R})$ where G represents the event that the image is generated and R represents the event that the image is real for each participant p. We evaluate this hypothesis using a two-sample t test with significance level α=.10, where the first sample is the participant's binary predictions for generated images, and the second is their predictions for real images. We define the null hypothesis $\mathbb{P}(\text{random } p \in P \text{ guesses real} \mid \text{G}) = \mathbb{P}(\text{random } p \in P \text{ guesses real} \mid \text{R})$ to evaluate VTTs for multiple participants P. We evaluate this hypothesis via a two-sample t test with significance level α=.10, where the first sample is the FPR and the second is the true positive rate of each participant.

2.3 Fréchet Distances

Quantitative evaluation of synthetic image quality was performed by calculating the FD $d(\Sigma_1, \Sigma_2, \mu_1, \mu_2)^2 = |\mu_1 - \mu_2|^2 + \text{tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1 \Sigma_2)^{\frac{1}{2}})$ [23] between two multivariate Gaussians (Σ_R, μ_R) and (Σ_G, μ_G) fitted to real and generated features extracted from the penultimate layer of eleven backbone networks: InceptionV3 [24], ResNet50 [25], InceptionResNetV2 [26], and DenseNet121 [27] each trained separately on both ImageNet [28] and RadImageNet [6], along with SwAV [29], DINO [30], and a Swin Transformer [31] trained on ImageNet. The first four networks were included to compare all publicly available RadImageNet models to their ImageNet equivalents. SwAV and DINO were included to evaluate the impact of self-supervision, as self-supervised representations have demonstrated superior transferability to new domains [10] and richer embeddings on medical images [4]. Finally, a Swin Transformer [31] was included as transformers have been shown to create transferable and robust representations [32]. We are the first to use self-supervised and transformer architectures with FD for generative medical imaging evaluation. FDs were calculated between the entire real dataset and 50,000 generated images. ImageNet-based FDs were calculated with the StudioGAN repository [33]. Further implementation details are available on our GitHub.

As the scale of FDs varies substantially by feature extractor, relative FDs (rFDs) $\frac{d(\Sigma_R, \Sigma_G, \mu_R, \mu_G)^2}{d(\Sigma_{R_1}, \Sigma_{R_2}, \mu_{R_1}, \mu_{R_2})^2}$ were computed with a random split of the real features into two Gaussian distributions $(\Sigma_{R_1}, \mu_{R_1})$ and $(\Sigma_{R_2}, \mu_{R_2})$. Paired t tests with significance level α=0.05 were conducted on the FDs to benchmark the data augmentation techniques. The Pearson correlation coefficient (ρ) with significance level α=0.05 was used to quantify the correspondence between FDs and VTT metrics and the correspondence between individual FDs. We are the first to consider whether medical image-based FDs are correlated with human judgment.

3 Results

Table 1 summarizes the overall results of the VTTs, with detailed individual participant outcomes available on our GitHub. The rFDs based on ImageNet and RadImageNet are outlined in Tables 2 and 3, while the FDs can be found in Tables S1 and S2 in the supplementary material. Model rankings based on individual metrics are illustrated in Figure 1. Our analysis revealed consistent rankings among all ImageNet-based FDs, aligning closely with human judgment. In contrast, RadImageNet-based FDs exhibited volatility and diverged from human assessment. DiffAugment was the best-performing form of augmentation, generating hyper-realistic images on two datasets.

ImageNet extractors aligned with human judgment. ImageNet-based FDs were consistent with one another in ranking generative models, except for on the MSD dataset, where human rankings were also inconsistent (Figure 1). This consistency was reinforced by strong correlations between the FD derived from InceptionV3 and all other ImageNet-based FDs across all sixteen models ($.84 < \rho < .99, p < .001$). Furthermore, the ImageNet-based FD rankings aligned with expert judgment (Figure 1). On the ChestX-ray14 dataset, ImageNet-based FDs ranked generative models in the same order as the radiologists: DiffAugment, ADA, no augmentation, and APA. Particularly promising was the SwAV-based FD, which significantly correlated with human perception across all sixteen models ($\rho = .475$ with Diff, $p = .064$).

RadImageNet extractors were volatile. RadImageNet-based FDs produced inconsistent rankings that were contrary to expert judgment. Notably, on the SLIVER07 dataset, RadImageNet-based FDs ranked DiffAugment as one of the poorest-performing models. However, all measures of human judgment identified DiffAugment as the best-performing model (see Figure 1). This discrepancy is especially concerning considering RadImageNet's inclusion of approxi-

Table 1. VTT results. Columns 1 and 2 split the models by dataset and augmentation technique (Aug): no augmentation (None), ADA, APA, and DiffAugment (DiffAug). Columns 3 and 4 show average FPRs and FNRs, with FPRs near 50% implying random guessing. Column 5 provides t test p-values, which tested if participants selected "real" for real and generated images equally. Column 6 displays the average difference between mean Likert ratings for real and generated images (Diff), with negative values indicating that generated images were perceived as more realistic than real images. Column 7 presents KS test p-values, which tested if Likert ratings for real and generated images came from the same distribution. ↑ and ↓ denote preferable higher or lower values. Underlined boldface type represents the best performance per dataset. Gray boxes indicate failure to reject the null hypothesis, suggesting that participants viewed real and generated images as equivalent. † indicates decreased performance compared to no augmentation.

Dataset	**Aug**	**FPR** [%] ↑	**FNR** [%] ↑	t **Test**	**Diff** ↓	**KS Test**
ChestXray-14	None	**48**	**58**	p=.497	0.12	p=.869
	ADA	32†	47†	p=.340	0.28†	p=.549
	APA	34†	56†	p=.082	0.24†	p=.717
	DiffAug	**48**	**58**	p=.616	**-0.16**	p=.967
SLIVER07	None	20	**34**	p=.424	0.68	p<.001
	ADA	24	30†	p=.748	0.52	p=.001
	APA	10†	28†	p=.232	0.82†	p<.001
	DiffAug	**34**	30†	p=.825	**0.22**	p=.717
MSD	None	58	48	p=.543	0.08	p>.999
	ADA	**66**	48	p=.217	-0.04	p>.999
	APA	46†	38†	p=.587	0.04	p>.999
	DiffAug	50†	**54**	p=.812	**-0.08**	p>.999
ACDC	None	34	22	p=.470	0.52	p=.022
	ADA	38	**30**	p=.653	0.38	p=.112
	APA	28†	22	p=.707	0.46	p=.003
	DiffAug	**44**	16†	p=.015	**0.28**	p=.112

mately 300,000 CT scans. On the ChestX-ray14 dataset, the FD derived from a RadImageNet-trained InceptionV3 network ranked the model without augmentation as the best performing. In contrast, a thoracic radiologist observed that both the APA and no augmentation models generated multiple radiographs with obviously distorted anatomy. Conversely, the weaknesses of the DiffAugment and ADA models were more subtle, with mistakes in support devices and central lines.

APA and ADA demonstrated varied performance. Although APA was designed to enhance image quality in limited data domains such as medical imaging, it unexpectedly reduced the perceptual quality of the generated images (t test on FPRs, $p = .012$), leading to an 18% reduction in the FPR on average. While ADA outperformed APA (t test on FDs, $p = .050$), it did not significantly affect participants' ability to differentiate real from generated images (t test on FPRs, $p > .999$). Despite both techniques underperforming in the VTTs, they

Table 2. ImageNet-based rFDs. Columns 1 and 2 split the models by dataset and augmentation technique (Aug): no augmentation (None), ADA, APA, and DiffAugment (DiffAug). Columns 3-9 show rFDs computed using seven ImageNet-based extractors: InceptionV3 (Incept), ResNet50 (Res), InceptionResNetV2 (IRV2), DenseNet121 (Dense), SwAV, DINO, and Swin Transformer (Swin). ↓ denotes lower values are preferable. Underlined boldface type indicates the best performance per dataset. † denotes decreased performance compared to no augmentation.

Dataset	**Aug**	**Relative Fréchet Distances** (ImageNet) ↓						
		Incept	Res	IRV2	Dense	SwAV	DINO	Swin
ChestXray-14	None	12.53	279.00	701.00	20.80	53.50	60.43	34.00
	ADA	8.90	237.00	576.00	15.55	33.00	37.81	26.36
	APA	17.58†	334.00†	1004.50†	39.85†	66.00†	82.23†	54.21†
	DiffAug	**<u>7.68</u>**	**<u>146.00</u>**	**<u>441.00</u>**	**<u>13.25</u>**	**<u>25.00</u>**	**<u>34.51</u>**	**<u>22.79</u>**
SLIVER07	None	1.48	7.90	12.98	2.59	8.28	6.12	6.07
	ADA	1.24	7.35	11.71	1.95	6.86	4.57	6.22†
	APA	1.37	7.33	11.96	2.36	7.79	5.59	5.43
	DiffAug	**<u>0.78</u>**	**<u>3.25</u>**	**<u>5.99</u>**	**<u>1.24</u>**	**<u>5.26</u>**	**<u>3.07</u>**	**<u>4.77</u>**
MSD	None	37.32	63.13	61.18	170.38	142.50	**<u>108.39</u>**	504.47
	ADA	**<u>36.84</u>**	**<u>62.50</u>**	**<u>58.88</u>**	**<u>141.63</u>**	305.00†	121.90†	308.59
	APA	43.63†	70.00†	81.76†	145.13	**<u>122.50</u>**	126.47†	196.65
	DiffAug	46.32†	125.50†	79.88†	170.38	825.00†	138.11†	**<u>175.12</u>**
ACDC	None	49.67	86.48	121.14	87.46	118.00	140.15	111.07
	ADA	20.99	31.66	49.94	35.95	76.40	65.52	61.49
	APA	31.15	54.35	76.47	56.68	90.60	87.69	72.10
	DiffAug	**<u>15.87</u>**	**<u>23.58</u>**	**<u>40.60</u>**	**<u>27.20</u>**	**<u>71.00</u>**	**<u>50.47</u>**	**<u>47.23</u>**

improved the rFDs for the SLIVER07 (t tests, $p = .025$ ADA, $p = .016$ APA) and ACDC (t tests, $p = .003$ ADA, $p = .004$ APA) datasets.

DiffAugment created hyper-realistic images. DiffAugment outperformed the other augmentation techniques across all FDs (t tests, $p = .092$ ADA, $p = .059$ APA). DiffAugment was the only form of augmentation to significantly enhance perceptual quality (t test on Diff, $p = .001$), resulting in an 81% reduction in the average difference between mean Likert ratings. Participants rated images from DiffAugment-based models as more realistic than those from both the ChestX-ray14 and MSD datasets. Additionally, Likert ratings for real and generated images from all DiffAugment-based models did not differ significantly (KS test, $p = .793$), suggesting that participants perceived them as equivalent.

4 Discussion

RadImageNet-based FDs may have underperformed due to several factors. First, networks trained for disease detection place a greater emphasis on local regions than their ImageNet counterparts [8], likely affecting their ability to evaluate the quality of an entire image. Second, medical images are highly heterogeneous, including differences across modalities, acquisition protocols, patient populations,

Table 3. RadImageNet-based rFDs. Columns 1 and 2 split the models by dataset and augmentation technique (Aug): no augmentation (None), ADA, APA, and DiffAugment (DiffAug). Columns 3-6 display rFDs computed using four RadImageNet-based extractors: InceptionV3 (Incept), ResNet50 (Res), InceptionResNetV2 (IRV2), and DenseNet121 (Dense). ↓ denotes lower values are preferable. Underlined boldface type indicates the best performance per dataset. † denotes decreased performance compared to no augmentation.

		Relative Fréchet Distances (RadImageNet) ↓			
Dataset	**Aug**	Incept	Res	IRV2	Dense
ChestXray-14	None	<u>**140.00**</u>	75.00	<u>**80.00**</u>	40.00
	ADA	660.00†	135.00†	190.00†	80.00†
	APA	280.00†	65.00	<u>**80.00**</u>	80.00†
	DiffAug	280.00†	<u>**50.00**</u>	90.00†	<u>**30.00**</u>
SLIVER07	None	3.67	3.14	6.00	4.33
	ADA	<u>**1.89**</u>	<u>**1.86**</u>	3.75	<u>**2.33**</u>
	APA	2.22	<u>**1.86**</u>	<u>**3.00**</u>	2.67
	DiffAug	4.67†	3.29†	5.50	4.67†
MSD	None	53.00	32.50	<u>**32.50**</u>	<u>**40.00**</u>
	ADA	<u>**36.00**</u>	<u>**27.5**</u>	37.50†	60.00†
	APA	54.00†	32.50	40.00†	<u>**40.00**</u>
	DiffAug	1551.00†	1105.00†	350.00†	615.00†
ACDC	None	26.64	19.00	20.33	32.50
	ADA	<u>**10.18**</u>	9.25	<u>**9.67**</u>	13.00
	APA	14.09	<u>**8.75**</u>	11.67	17.50
	DiffAug	12.09	15.25	<u>**9.67**</u>	<u>**10.50**</u>

and image processing techniques. RadImageNet does not contain chest X-rays nor cine MRIs. Furthermore, it was collected from a single radiology facility [6], making it likely that the protocols, machinery, and patients populations differed from those of the SLIVER07 and MSD datasets.

5 Conclusion

Our study challenges prevailing assumptions by providing novel evidence that medical image-trained feature extractors do not inherently improve FDs for synthetic medical imaging evaluation; instead, they may compromise metric consistency and alignment with human judgment, even on in-domain data.

Acknowledgments. Research reported in this publication was supported in part by resources of the Image Guided Cancer Therapy Research Program (IGCT) at The University of Texas MD Anderson Cancer Center, a generous gift from the Apache Corporation, the National Institutes of Health/NCI under award number P30CA016672, and the Tumor Measurement Initiative through the MD Anderson Strategic Initiative Development Program (STRIDE). We thank the NIH Clinical Center for the ChestXray14 dataset, Dr. Rishi Agrawal and Dr. Carol Wu for their generative modeling feedback, Dr. Vikram Haheshri and Dr. Oleg Igoshin for the discussion that led to

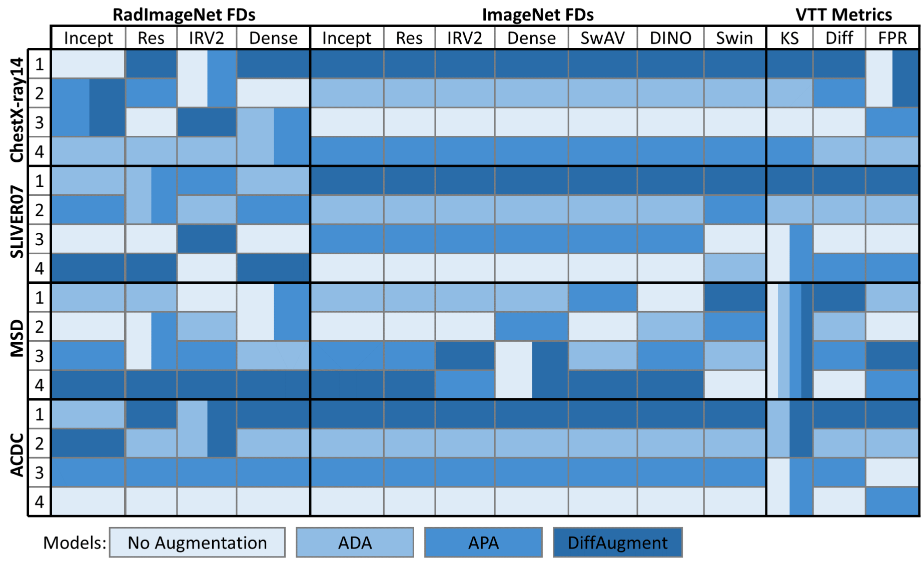

Fig. 1. Model rankings. Columns represent evaluation metrics. FDs are split by feature extraction dataset and architecture: InceptionV3 (Incept), ResNet50 (Res), Inception-ResNetV2 (IRV2), DenseNet121 (Dense), SwAV, DINO, and Swin Transformer (Swin). Metrics evaluating human perception via VTTs are KS test p-values (KS), average difference in mean Likert scores (Diff), and FPRs. Rows represent models trained with different augmentation techniques on the same dataset. Models are ranked 1-4 in descending order of performance and are differentiated by color. Vertical bars denote a shared rank.

the hypothesis testing contribution, and Erica Goodoff - Senior Scientific Editor in the Research Medical Library at MD Anderson for editing this article. GPT4 was used in the proofreading stage of this manuscript.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Guyon, I., et al. (eds.) NIPS. vol. 30. Curran Associates, Inc. (2017)
2. Woodland, M., et al.: Evaluating the performance of stylegan2-ada on medical images. In: Zhao, C., et al. (eds.) SASHIMI. pp. 142–153. Springer (2022). https://doi.org/10.1007/978-3-031-16980-9_14
3. Borji, A.: Pros and cons of gan evaluation measures. Comput. Vis. Image Underst. **179**, 41–65 (2019). https://doi.org/10.1016/j.cviu.2018.10.009
4. Truong, T., Mohammadi, S., Lenga, M.: How transferable are self-supervised features in medical image classification tasks? In: Jung, K., et al. (eds.) MLHC. vol. 158, pp. 54–74. PMLR (2021)

5. Kynkäänniemi, T., Karras, T., Aittala, M., Aila, T., Lehtinen, J.: The role of imagenet classes in fréchet inception distance. arXiv:2203.06026 (2023)
6. Mei, X., et al.: Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. Radiol.: Artif. Intell. **4**(5) (2022). https://doi.org/10.1148/ryai.210315
7. Osuala, R., et al.: medigan: a Python library of pretrained generative models for medical image synthesis. J. Med. Imaging **10**(6), 061403 (2023). https://doi.org/10.1117/1.JMI.10.6.061403
8. Anton, J., et al.: How well do self-supervised models transfer to medical imaging? J. Imaging **8**(12), 320 (2022). https://doi.org/10.3390/jimaging8120320
9. Morozov, S., Voynov, A., Babenko, A.: On self-supervised image representations for gan evaluation. In: ICLR (2020)
10. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Zabih, R., et al. (eds.) CVPR. IEEE (2020)
11. Chen, J., Wei, J., Li, R.: Targan: target-aware generative adversarial networks for multi-modality medical image translation. In: de Bruijne, M., et al. (eds.) MICCAI. pp. 24–33. Springer (2021). https://doi.org/10.1007/978-3-030-87231-1_3
12. Jung, E., Luna, M., Park, S.H.: Conditional gan with an attention-based generator and a 3d discriminator for 3d medical image generation. In: de Bruijne, M., et al. (eds.) MICCAI. pp. 318–328. Springer (2021). https://doi.org/10.1007/978-3-030-87231-1_31
13. Tronchin, L., Sicilia, R., Cordelli, E., Ramella, S., Soda, P.: Evaluating gans in medical imaging. In: Engelhardt, S., et al. (eds.) DGM4MICCAI, DALI. pp. 112–121. Springer (2021). https://doi.org/10.1007/978-3-030-88210-5_10
14. Heimann, T., et al.: Comparison and evaluation of methods for liver segmentation from ct datasets. IEEE Trans. Med. Imaging **28**(8), 1251–1265 (2009). https://doi.org/10.1109/TMI.2009.2013851
15. Wang, X., et al.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Chellappa, R., et al. (eds.) CVPR. IEEE (2017)
16. Antonelli, M., et al.: The medical segmentation decathlon. Nat. Commun. **13**(1), 4128 (2022). https://doi.org/10.1038/s41467-022-30695-9
17. Simpson, A.L., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv:1902.09063 (2019)
18. Bernard, O., et al.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE Trans. Med. Imaging **37**(11), 2514–2525 (2018). https://doi.org/10.1109/TMI.2018.2837502
19. Karras, T., et al.: Analyzing and improving the image quality of stylegan. In: Zabih, R., et al. (eds.) CVPR. IEEE (2020)
20. Karras, T., et al.: Training generative adversarial networks with limited data. In: Larochelle, H., et al. (eds.) NeurIPS. vol. 33, pp. 12104–12114. Curran Associates, Inc. (2020)
21. Zhao, S., Liu, Z., Lin, J., Zhu, J.Y., Han, S.: Differentiable augmentation for data-efficient gan training. In: Larochelle, H., et al. (eds.) NeurIPS. vol. 33, pp. 7559–7570. Curran Associates, Inc. (2020)
22. Jiang, L., Dai, B., Wu, W., Loy, C.C.: Deceive d: Adaptive pseudo augmentation for gan training with limited data. In: Ranzato, M. (ed.) NeurIPS. vol. 34, pp. 21655–21667. Curran Associates, Inc. (2021)
23. Dowson, D., Landau, B.: The fréchet distance between multivariate normal distributions. J. Multivar. Anal. **12**(3), 450–455 (1982). https://doi.org/10.1016/0047-259X(82)90077-X

24. Szegedy, C., et al.: Going deeper with convolutions. In: Bischof, H., et al. (eds.) CVPR. IEEE (2015)
25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Zabih, R., et al. (eds.) CVPR. IEEE (2016)
26. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning. vol. 31 (2017). https://doi.org/10.1609/aaai.v31i1.11231
27. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR. IEEE (2017)
28. Deng, J., et al.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. IEEE (2009). https://doi.org/10.1109/CVPR.2009.5206848
29. Caron, M., et al.: Unsupervised learning of visual features by contrasting cluster assignments. In: NeurIPS. vol. 33, pp. 9912–9924. Curran Associates, Inc. (2020)
30. Caron, M., et al.: Emerging properties in self-supervised vision transformers. In: Berg, T., et al. (eds.) ICCV. pp. 9650–9660. IEEE (2021)
31. Li, Z., Wang, Y., Yu, J.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Berg, T., et al. (eds.) ICCV. pp. 10012–10022. IEEE (2021)
32. Zhou, H.Y., Lu, C., Yang, S., Yu, Y.: Convnets vs. transformers: Whose visual representations are more transferable? In: Vandenhende, S., et al. (eds.) ICCV Workshops. pp. 2230–2238. IEEE (2021)
33. Kang, M., Shim, W., Cho, M., Park, J.: Studiogan: A taxonomy and benchmark of gans for image synthesis. Trans. Pattern Anal. Mach. Intell. **45**(12), 15725–15742 (2023). https://doi.org/10.1109/TPAMI.2023.3306436