F?D: On understanding the role of deep feature spaces on face generation evaluation

Krish Kabra
Rice University
kk80@rice.edu

Guha Balakrishnan
Rice University
guha@rice.edu

Abstract

Perceptual metrics, like the Fréchet Inception Distance (FID), are widely used to assess the similarity between synthetically generated and ground truth (real) images. The key idea behind these metrics is to compute errors in a deep feature space that captures perceptually and semantically rich image features. Despite their popularity, the effect that different deep features and their design choices have on a perceptual metric has not been well studied. In this work, we perform a causal analysis linking differences in semantic attributes and distortions between face image distributions to Fréchet distances (FD) using several popular deep feature spaces. A key component of our analysis is the creation of synthetic counterfactual faces using deep face generators. Our experiments show that the FD is heavily influenced by its feature space's training dataset and objective function. For example, FD using features extracted from ImageNet-trained models heavily emphasizes hats over regions like the eyes and mouth. Moreover, FD using features from a face gender classifier emphasizes hair length more than distances in an identity (recognition) feature space.

1. Introduction

Rapid advances in generative image models such as variational autoencoders (VAEs) [27, 34], generative adversarial networks (GANs) [8, 14, 20, 21], and diffusion models [12, 17, 28, 30], point to a future where synthetic images play a significant role in society [15, 23, 36]. Therefore, it is crucial to continuously assess and improve how we evaluate the performances of these generative models [6]. In particular, synthesis evaluation metrics should capture several factors, including correlation to human perception, robustness to insignificant variations and noise, and sensitivity to domain-specific semantics.

The gold standard in evaluating image generation quality is human annotation [40], which can provide nuanced and interpretable perceptual feedback, but comes at the cost of money and time. The current standards in automated evaluation are deep perceptual metrics, which embed images into lower-dimensional representations derived from the final layers of deep neural networks and compute distances between images [2, 39]. In particular, the Fréchet Inception Distance (FID) [16] is currently the *de facto* image generation evaluation metric. FID calculates the Fréchet distance (FD) [13] between two multivariate Gaussians fitted to representations extracted from the InceptionV3 [31]) trained on ImageNet [10] for real and generated images:

$$\begin{aligned}\text{FD}(\mu_{\mathbf{1}}, \mathbf{\Sigma_1}, \mu_{\mathbf{2}}, \mathbf{\Sigma_2}) = ||\mu_{\mathbf{1}} - \mu_{\mathbf{2}}||_2^2 \\ + \text{Tr}\Big(\mathbf{\Sigma_1} + \mathbf{\Sigma_2} - 2(\mathbf{\Sigma_1}\mathbf{\Sigma_2})^{\frac{1}{2}}\Big),\end{aligned} \quad (1)$$

where $(\mu_{\mathbf{1}}, \mathbf{\Sigma_1})$ and $(\mu_{\mathbf{2}}, \mathbf{\Sigma_2})$ are the sample means and covariances of the real and generated image set embeddings, and $\text{Tr}(\cdot)$ is the matrix trace.

While these metrics have been shown to correlate better with human evaluation than classical metrics computed directly from image pixels (e.g. average log-likelihood) [32, 38], unfortunately, the complexity of deep feature spaces also makes them opaque and hard to interpret. Given that deep generative models are now typically competing with each other for less than 5 FID points, it is unclear what such differences mean semantically. When evaluating face generators, answering questions such as "What effect does an imbalanced generation of skin tones have on FID?" or, "What is the effect of consistent distortion of the eyes on FID?" is crucial in helping engineers better understand their evaluation metric, which ultimately will enable them to mitigate biases inherent in their models and improve generation quality.

In this work, we propose a strategy to causally evaluate the effect variations in domain-specific characteristics have on a deep perceptual evaluation metric using synthetic data. We focus on face generation, the most popular domain for image synthesis studies with many important societal implications in applications like face analysis/recognition [3], deepfakes [33], virtual avatars [1, 25], and even healthcare

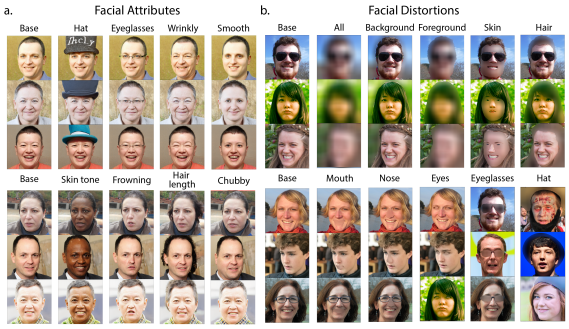

Figure 1. **Sample images in our proposed counterfactual face dataset**. **a.** The base column shows synthetically generated base faces with predefined "neutral" characteristics. We manipulate each base face along different attribute of interest, shown in the remaining columns. **b.** The base column shows real faces from FFHQ. Faces in the remaining columns are distorted versions of the base faces with blur added to specific semantic regions. Note that the examples shown for eyes, eyeglasses, and hat are different because the corresponding base faces do not contain those regions.

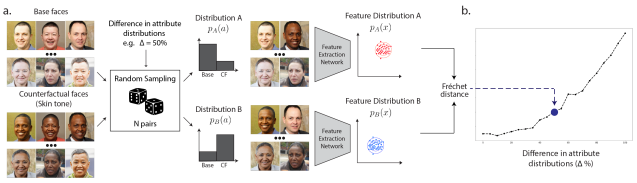

Figure 2. **Method overview for measuring causal effects of semantic attribute differences on Fréchet distance. a.** We generate a set of base and counterfactual (CF) face pairs for an attribute or semantic distortion (for this figure, we use skin tone as an example). Given a difference in proportion for this attribute between two distributions ($\Delta \in [0\%, 100\%]$) and the number of faces per set (N), we construct two image sets A and B by randomly assigning base and CF faces to them such that this difference is achieved. We then extract the features for each image set using a pre-trained deep model (e.g., Inception, CLIP), and compute the FD between the two feature distributions. **b.** By creating set pairs for $\Delta \in [0\%, 100\%]$, we can generate a curve that summarizes the causal effect of a difference in attribute proportions on FD computed in a feature space.

[37]. We consider two types of variations: differences in semantic attributes (e.g., hats, skin tone, hair length) and distortions (blur) to semantic regions (e.g, nose, eyes, mouth). We perform causal studies using experimental interventions that manipulate a single characteristic of an image at a time, enabling us to create synthetic counterfactuals. For semantic attribute interventions, we use deep face generators to construct a dataset of synthetic face pairs that differ (approximately) by only a single feature of interest. For distortions, we apply blur to semantic facial regions inferred by a face segmentation model. Using this synthetic data, we measure the causal effects the studied characteristic variations have on the FD in six different deep representations. The results demonstrate that deep feature spaces have significant and unique biases over in-domain attributes due to both training data and objective functions. These biases should be understood by researchers during synthesis evaluation.

2. Methods

For a given deep feature space, our goal is to quantify the sensitivity of an evaluation metric to image characteristics. In our experiments, we focus on face images and FD, and so we describe our methods here in that context. We form two questions for a given feature space: (1) How do differences between the semantic attribute distributions of two face image sets quantitatively affect FD? (2) How do distortions localized to a semantic region of a face quantitatively affect FD? These questions align with two broad image characteristics that a generative model must capture: (1) semantic attributes for the domain, and (2) realistic details.

Answering these questions requires causal reasoning, and ideally a *counterfactual* dataset consisting of pairs of faces that are identical except for a difference along one characteristic (i.e., semantic attribute or distortion). Real-face datasets contain significant attribute correlations [3] and are therefore not appropriate. Instead, we propose a synthetic approach. In the following sections, we outline the proposed methodology to construct synthetic data to answer each question. Example images synthesized by the proposed methods are shown in Fig. 1.

2.1. Measuring the effect of semantic attribute differences on Fréchet distance

Consider two image sets A and B with feature distributions $p_A(x)$ and $p_B(x)$, where $x \in R^D$ is a feature space of an image. Furthermore, assume that A and B are identical except for a difference in their distributions over one semantic binary attribute with value $a \in \{0, 1\}$, which we denote by $p_A(a)$ and $p_B(a)$. Our goal is to quantify how the difference in attribute proportions between $p_A(a)$ and $p_B(a)$ (ranging from 0% when identical to 100% when completely dissimilar), affects $\text{FD}(\mu_A, \Sigma_A, \mu_B, \Sigma_B)$, the FD between $p_A(x)$ and $p_B(x)$. Fig. 2 describes our analysis methodology to do so. We construct multiple sets of nearly identical faces using deep generative models (described below), each consisting of different proportions of values to a, and compute FD between the pairs to yield a curve summarizing causal effects (see Fig. 2-right, and Fig. 3).

This analysis requires the creation of sets A and B, counterfactual face sets that differ based on only a. We use a two-step process to create this data synthetically. First, we synthesize a set of *base faces* that exhibit predefined uniform characteristics of light skin tones and short hair, and

no: facial hair, make-up, frowning expressions, hats, or eyeglasses. To do this, we obtained the face generation models of a previous facial causal benchmarking study [3] based on StyleGAN2 [21] trained on the Flickr-Faces-HQ (FFHQ) dataset [20] and orthogonalized linear latent space traversals (OLLT). We filter these faces via human evaluations to ensure they meet the defined criteria. In our experiments, we used a total of 1427 filtered base faces. In the second step, we synthesize counterfactual pairs from the base faces for each attribute a. In our experiments, we analyzed 8 binary attributes corresponding to various facial semantics including geometry, skin tone, skin texture, hair length, and accessories. The attributes analyzed are shown by the columns in Fig. 1a. We utilize one of three different image manipulation methods based on the attribute type: (1) OLLT, (2) StyleCLIP [24], and (3) image inpainting with Stable Diffusion [35]. We choose the best method for each attribute based on a qualitative assessment of how well each method can manipulate the attribute while holding others constant. We show some example attribute counterfactuals in Fig. 1a. We provide a complete account of models, experimental parameters, and details used to create the synthetic dataset in Supplementary.

2.2. Measuring the effect of blurring semantic regions on Fréchet distance

The purpose of this analysis is to understand how a systematic distortion outputted by a face generator for a specific semantic region impacts FD. In our experiments, we focused on heavy blur, though many others may also be explored. For each region, we use real FFHQ face images that contain that region (accessories like hats and eyeglasses are not in every image) as one distribution (set A), and apply Gaussian blur to these images *only in that region* using segmentation masks obtained from a public face segmentation model[1] (set B). We considered 9 regions in our experiments, as shown in Fig. 1b. For our analysis, we simply report FD with respect to distorting each semantic region (see Fig. 4).

3. Experiments

We conduct our analyses using six deep feature spaces with publicly available parameters: (1) Inception V3 model trained on the ILSVRC-2012 (ImageNet) dataset for classification [31], (2) CLIP (ViT-B/32) model trained on large-scale dataset of image-text pairs using a contrastive loss [26], (3) SwAV (ResNet-50) model trained on the ImageNet in self-supervised scheme [7], (4) FairFace (ResNet-34) model trained on FairFace for race, age, and gender classification [19], (5) SwAV-FFHQ (ResNet-50) model trained on FFHQ in self-supervised scheme [4, 7], and (6) Identity ArcFace (ResNet-34) model trained on Glint360k for facial recognition [11].

[1] https://github.com/zllrunning/face-parsing.PyTorch

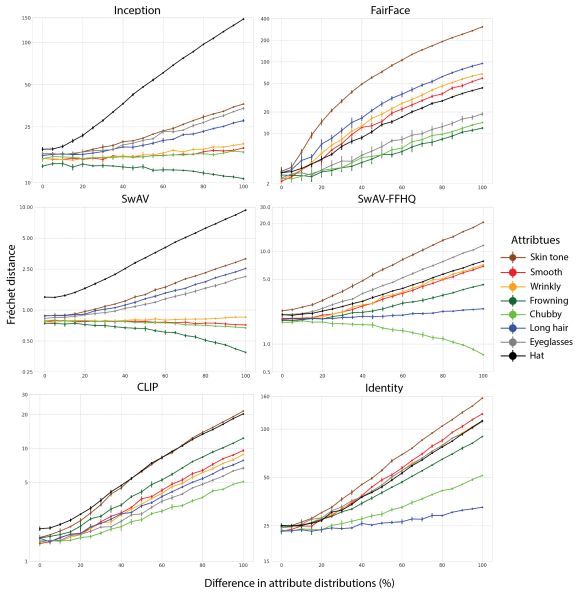

Figure 3. **Results for causal sensitivity analysis of Fréchet distances (FD) in different feature spaces with respect to semantic attributes.** For each percentage difference in attribute proportion (i.e. point along the x-axis), we sample 10 random draws of 1000 counterfactual face pairs to construct face sets, from which the FD in a feature space is computed, shown by the y-axis in log scale. Each feature space under- or over-emphasizes certain attributes based on its training dataset and objective functions.

We present sensitivity analyses of FD with respect to facial attribute proportions in Fig. 3. Our total dataset used for this analysis contains 1427 counterfactual face pairs. For each percentage difference in attribute proportion (i.e. point along the x-axis), we sample 10 random draws of 1000 pairs to construct face sets, from which the FD in a feature space is computed. The points and error bars shown in the plots correspond to the mean and standard deviation respectively. A direct comparison of FD values across feature spaces is not meaningful, as the scale of distances vary across features. However, the difference in trends between the attribute curves may be compared across plots. For example, Inception and SwAV clearly emphasize hats with respect to other feature spaces, while FairFace and SwAV-FFHQ emphasize skin tone.

We present sensitivity analyses of FD with respect to localized distortions in Fig. 4. To compare FD across different feature spaces, we normalize distances by dividing them by the distance between the original and entirely blurry images ("all" category in Fig. 1) in that feature space.

4. Discussion

Feature spaces learned using ImageNet underemphasize important facial semantics regardless of

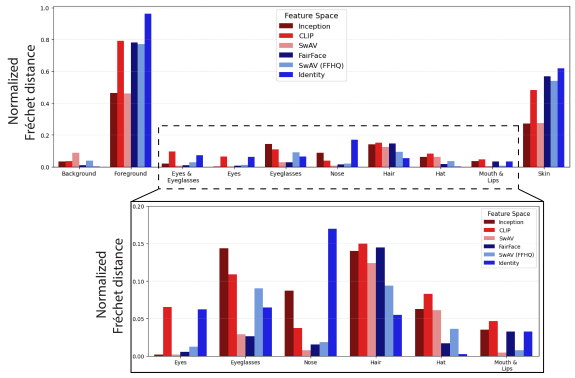

Figure 4. **Results for semantic region distortion (blur) analysis.** (Top) Bar plot comparing the normalized FD per semantic region blur. We normalize distances in each feature space by dividing by the distance between the original and fully blurred image set ("All" in Fig. 1) in that space. (Bottom) Zoomed-in plot to clearly visualize results for semantic regions that occupy less than 25% of the face image on average.

the training objective. Fig. 3 illustrates that FD in feature spaces learned using ImageNet (Inception and SwAV) are most sensitive to differences in the proportion of hats, consistent with findings from Kynkäänniemi *et al.* [22]. However, interestingly, the FD computed using SwAV features is also sensitive to hats, even though that model is not explicitly trained to classify ImageNet classes. This is reasonable since self-supervised learning is known to be an effective pretraining strategy for ImageNet classification. The plots also illustrate that FD computed using ImageNet-learned spaces are highly insensitive to distributional differences in skin texture ("wrinkly" and "smooth"), geometry ("chubby"), and expression ("frowning"). A deeper investigation (see Supplementary) reveals a subtle interplay between the mean and trace terms of the FD in Eq. (1). As the two distributions become more skewed in our sensitivity analyses (towards 0 or 100 % in Fig. 3), the distribution means become more dissimilar, but their variances also decrease and reduce the trace term. This suggests another challenge in using FD alone: they can obfuscate differences in distribution modes versus distribution shapes.

Fig. 4 shows that FD computed using Inception and SwAV spaces are insensitive to the blurring of the eyes, and SwAV is insensitive to the blurring of the nose and mouth. This shows that systematic degradations to the eyes, nose, or mouth, will not impact the FD in ImageNet-based feature spaces. Generative model designers should pay extra attention to these semantic "blind spots."

The training objective influences which facial semantics are emphasized by a deep feature space. Fig. 3 shows that while in-domain feature spaces (FairFace, SwAV-FFHQ, Identity) are all highly sensitive to differences in skin tone, skin texture, and facial accessories, there do exist several notable dissimilarities. For example, FairFace is far more sensitive to hair length, compared to SwAV-FFHQ and Identity. This is further supported by the relatively small effect that blurring the hair has on SwAV-FFHQ and Identity compared to FairFace. Another notable distinction is that both FairFace and SwAV-FFHQ fail to capture distortions localized to the eyes, nose, mouth, and lips, whereas Identity does. We speculate that these differences are a consequence of the feature spaces capturing semantic characteristics that pertain most to the objective function used during training. FairFace is trained to classify perceived gender, which is correlated with hair length. On the other hand, Identity is trained to match faces corresponding to the same person, which should be invariant to hairstyle and hair length. SwAV is trained to match cropped views of an image, for which hair length is likely not a robust feature. Therefore, we suggest that generative model designers should not naïvely expect in-domain feature spaces to be sensitive to all domain-specific semantics. Rather, we advocate carefully considering how the training objective may influence features, and empirically investigating these sensitivities.

Image-language models trained on massive general datasets capture many important semantic characteristics of faces. The sensitivity analyses for both semantic attributes and distortions show the CLIP features are sensitive to all studied characteristics. In particular, CLIP provides a significant FD for all distorted facial region irrespective of the size of the region in pixels. This is likely because of two reasons: (1) CLIP is trained on a massive dataset, and (2) text provides a rich source of information on perceptual features to the image encoder that cannot have otherwise been learned using classical supervision. Based on these results, we encourage generative model designers to move away from perceptual features extracted from models trained on ImageNet (Inception, VGG [29], SwAV) and use large image-language models like CLIP.

4.1. Limitations

Our causal analysis of semantic attributes assumes perfectly counterfactual face pairs. However, it is difficult to isolate one attribute from others when working with deep generators due to the correlations the generator learns from its training distribution. Nevertheless, in general, such correlations are known to be even more dramatic in real datasets [3], which makes synthetic generation a more attractive option for such analysis. Another limitation includes the sample size of 1000 images per set used in our semantic attribute analysis, which results in biased FD estimates [5, 9, 18]. However, given that sample size was consistent throughout the experiment, the trend and shapes of curves shown in Fig. 3 are accurate.

References

[1] Rameen Abdal, Hsin-Ying Lee, Peihao Zhu, Menglei Chai, Aliaksandr Siarohin, Peter Wonka, and Sergey Tulyakov. 3davatargan: Bridging domains for personalized editable avatars, 2023. 1

[2] Dan Amir and Yair Weiss. Understanding and simplifying perceptual distances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12226–12235, 2021. 1

[3] Guha Balakrishnan, Yuanjun Xiong, Wei Xia, and Pietro Perona. Towards Causal Benchmarking of Biasin Face Analysis Algorithms. In *Deep Learning-Based Face Analytics*, pages 327–359. Springer International Publishing, Cham, 2021. 1, 2, 3, 4

[4] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-Efficient Semantic Segmentation with Diffusion Models. In *International Conference on Learning Representations*, 2022. 3

[5] Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2023. 4

[6] Ryan Burnell, Wout Schellaert, John Burden, Tomer D. Ullman, Fernando Martinez-Plumed, Joshua B. Tenenbaum, Danaja Rutar, Lucy G. Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, Douwe Kiela, Murray Shanahan, Ellen M. Voorhees, Anthony G. Cohn, Joel Z. Leibo, and Jose Hernandez-Orallo. Rethink reporting of evaluation results in AI. *Science*, 380(6641):136–138, 2023. 1

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *Advances in Neural Information Processing Systems*, pages 9912–9924. Curran Associates, Inc., 2020. 3

[8] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient Geometry-Aware 3D Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 1

[9] Min Jin Chong and David Forsyth. Effectively Unbiased FID and Inception Score and Where to Find Them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6070–6079, 2020. 4

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 3

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, pages 8780–8794. Curran Associates, Inc., 2021. 1

[13] D. C Dowson and B. V Landau. The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982. 1

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, 63(11):139–144, 2020. 1

[15] Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3313–3332, 2023. 1

[16] 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*. Curran Associates, Inc., 2017. 1

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 1

[18] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. *arXiv preprint arXiv:2401.09603*, 2023. 4

[19] Kimmo Kärkkäinen and Jungseock Joo. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1547–1557, 2021. 3

[20] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 1, 3

[21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 1, 3

[22] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The Role of ImageNet Classes in Fréchet Inception Distance. In *The Eleventh International Conference on Learning Representations*, 2023. 4

[23] Daniel McDuff, Theodore Curran, and Achuta Kadambi. Synthetic data in healthcare, 2023. 1

[24] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 3

[25] Justin N. M. Pinkney and Doron Adler. Resolution dependent gan interpolation for controllable image synthesis between domains, 2020. 1

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen

Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3

[27] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating Diverse High-Fidelity Images with VQ-VAE-2. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 1

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 4

[30] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum Likelihood Training of Score-Based Diffusion Models. In *Advances in Neural Information Processing Systems*, 2021. 1

[31] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. 1, 3

[32] Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models, 2016. 1

[33] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A Survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020. 1

[34] Arash Vahdat and Jan Kautz. NVAE: A Deep Hierarchical Variational Autoencoder. In *Advances in Neural Information Processing Systems*, pages 19667–19679. Curran Associates, Inc., 2020. 1

[35] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models, 2023. 3

[36] Tonghe Wang, Yang Lei, Yabo Fu, Jacob F. Wynne, Walter J. Curran, Tian Liu, and Xiaofeng Yang. A review on medical imaging synthesis using deep learning and its clinical applications. *Journal of Applied Clinical Medical Physics*, 22(1): 11–36, 2021. 1

[37] Zhen Wang, Yunhao Ba, Pradyumna Chari, Oyku Deniz Bozkurt, Gianna Brown, Parth Patwa, Niranjan Vaddi, Laleh Jalilian, and Achuta Kadambi. Synthetic Generation of Face Videos With Plethysmograph Physiology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20587–20596, 2022. 2

[38] Qiantong Xu, Gao Huang, Yang Yuan, Chuan Guo, Yu Sun, Felix Wu, and Kilian Weinberger. An empirical study on evaluation metrics of generative adversarial networks. *arXiv preprint arXiv:1806.07755*, 2018. 1

[39] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 1

[40] Sharon Zhou, Mitchell Gordon, Ranjay Krishna, Austin Narcomey, Li F Fei-Fei, and Michael Bernstein. HYPE: A Benchmark for Human eYe Perceptual Evaluation of Generative Models. In *Advances in Neural Information Processing Systems*, 2019. 1