Beyond FID: Human Perceptual Judgments Reveal Systematic Blind Spots in GAN Face Evaluation

B. Nierula[1], A. Melnik[1], F. Barthel[1], A. Brama, A. Hilsmann[1], P. Eisert[1,2], V. V. Nikulin[3], M. Gaebler[3], F. Klotzsche[3], Y. Chen[3], T. Stephani[3,4], S. Bosse[1]

[1]Fraunhofer Heinrich-Hertz-Institute, Berlin, Germany; [2]Humboldt University of Berlin, Berlin, Germany; [3]Max-Planck-Institute for Human Cognitive and Brain Sciences, Leipzig, Germany; [4]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, the Netherlands

Abstract

Generative Adversarial Networks (GANs) can synthesize highly realistic facial images from random noise vectors. The Fréchet Inception Distance (FID) is widely used as a standard metric to automatically evaluate the quality of GAN-generated images. However, it remains unclear to what extent this statistical measure reflects human perceptual judgments, which ultimately define image realism in practical applications. To address this, we conducted a psychophysical study in which participants (n = 20) performed a two-alternative forced-choice task, assessing actual photographs and GAN-generated images as real or fake. We show that while FID provides a reliable global ordering of image quality, it systematically fails for localized semantic artifacts (e.g., eyewear and skin texture) that disproportionately affect human realness judgments. This demonstrates that FID and human perception are not merely noisy versions of the same signal, but that FID has systematic blind spots for localized semantic artifacts that disproportionately drive human realism judgments.

CCS Concepts

• ***Theory of computation*** → ***Models of computation;*** *Interactive computation;*

1. Introduction

Generative adversarial networks (GANs) [GPAM*14] synthesize human faces by learning a mapping from random latent vectors to images through an adversarial game between two deep neural networks: a generator that produces candidate faces and a discriminator that attempts to distinguish real faces from generated ones, driving the generator to model the underlying face distribution. During training, both networks are optimized with opposing objectives so that, at equilibrium, the generator's outputs become statistically indistinguishable from real face images (for review see [KST*23]).

The Fréchet Inception Distance (FID) [Fré57, DL82] was proposed by Heusel et al. as a quantitative measure for assessing the quality of GAN-generated images [HRU*17]. FID extracts feature embeddings from datasets of real and generated images using the Inception-v3 network [SVI*16]. Assuming that these embeddings follow multivariate normal distributions, FID quantifies the distance between real and generated image distributions by comparing their respective means and covariances. Lower FID values indicate higher statistical similarity between the distributions, suggesting that the generated images more closely resemble real samples. Owing to its simplicity and empirical usefulness, FID has become a widely adopted metric for evaluating image quality [KST*23] and comparing GAN models [LKM*18].

FID is sensitive to a range of image distortions, including Gaussian noise, Gaussian blur, implanted black rectangles, swirled images, salt and pepper noise, and dataset contamination [HRU*17], and performs well in terms of discriminability, robustness, and computational efficiency [Bor19]. However, the relationship between FID scores and human perceptual judgments remains controversial. While the original FID paper reports consistency with human judgments [HRU*17], subsequent studies show that FID can diverge substantially from human assessment [Bor19, LWLZ18, JRV*24]. A critical limitation of previous evaluations is the considerable variability in human perceptual judgments, both across and within individual observers [MBPW17]. Nevertheless, studies comparing FID to human perception often do not report their number of participants or rely on very few (1-3) raters [LWLZ18, JRV*24, HRU*17]. As a result, it is often unclear whether observed discrepancies reflect genuine limitations of FID and/or merely the inherent variability in human responses.

To address this, we systematically investigated the relationship between FID scores and human perceptual judgments by collecting realness ratings from 20 participants across images spanning different FID quality levels. Our contributions are: (1) a characterization of the relationship between FID values and human realness judgments across multiple raters; (2) an analysis of how FID values relate to judgment confidence; (3) an openly available data set to the community to further validate VID ratings.

In the following, we review related work on GAN evaluation metrics (Section 2), describe our experimental methods (Section 3),

DOI: 10.2312/egs.20261013

characterize the FID-human relationship (Section 4), and discuss implications for evaluating generative models (Section 5).

2. Related Work

To evaluate the quality of GAN-generated images, the FID [HRU*17] has become the most widely used metric in the field. Other measures such as the average log-likelihood [TGB*17], the inception score (IS) [SGZ*16], the maximum mean discrepancy (MMD) [GBR*12], HYPE [ZGK*19], CMMD [JRV*24] or perceptually grounded metrics from deep features [ZIE*18] are used less frequently (see also [Bor19] for an overview of GAN evaluation metrics).

FID performs well in terms of discriminability, robustness, and computational efficiency [Bor19]; however, it also suffers from several limitations. A common critique is that it assumes normally distributed feature embeddings [Bor19, JRV*24] and that it relies on Inception-v3 embeddings trained on a limited dataset, which does not adequately represent the diverse content produced by modern generative models [JRV*24]. The metric is also sensitive to low-level image processing operations, such as compression and resizing, and fails to capture certain complex image distortions. While the original authors point out its consistency with human judgments [HRU*17], subsequent studies have shown that it can contradict human perceptual assessment [MBPW17, JRV*24]. These limitations indicate that FID alone is insufficient for a comprehensive evaluation of generative models.

In this work, we relate FID to the gold standard of human judgments by examining how well the metric aligns with perceived image realness and observers' confidence in their ratings. Rather than asking whether FID agrees or disagrees with human judgments in absolute terms, our analysis treats FID as a conditionally valid metric: reliable for coarse realism estimation, but increasingly misaligned with perception as realism becomes dominated by localized semantic cues.

3. Methods

We conducted the following psychophysical experiment to measure how humans perceive the realness of GAN-generated face images across different quality levels.

Participants. 21 right-handed volunteers with normal or corrected-to-normal vision participated in the study. One participant was excluded from the study due to a misunderstanding of instructions, resulting in a final sample of 20 participants (10 females; $M = 27.55$ years, *range* = 22-39, $SD = 3.99$). Participants provided their informed consent prior to the experiment. The study was approved by the local ethics committee.

Stimuli. We synthesized realistic frontal-view human faces using the state-of-the-art CGS-GAN model [BMH*25], trained from scratch on the FFHQC dataset [BMH*25] (a cleaned version of FFHQ [KLA19], controlling for face occlusions). Training occurred at 512^2 resolution for 5M iterations (requiring approximately 3 days on 4×NVIDIA A100 GPUs). Since early training stages produced low-quality images, we began collecting samples only after the FID dropped below 25 and retained only improving checkpoints (strictly decreasing FID). We further applied the *truncation trick* [KLA19] for artifact reduction.

The resulting stimulus set comprised 1440 color images (720 real and 720 GAN-generated images). Synthetic images varied in image quality according to their FID scores from 4.682 (the lowest FID obtained by our model) to 20.840 (the threshold above which images are consistently perceived as fake, as determined in pilot studies). Figure 1 displays examples of a photograph and generated images at different FID levels.

Figure 1: *Examples for a actual photograph (real) and GAN-generated images with ascending FID values.*

Experiment. Participants completed a psychophysical experiment while seated in front of a computer screen. After a brief demographic questionnaire and on-screen instructions, each trial began with a 500 ms fixation cross, followed by a stimulus image presented for 350 ms. This duration supports stable conscious face perception [SBK*13] and neural processing of face realism [CSB*24, SZBK17]. Participants then judged whether the image was *real* or *fake* in a two-alternative forced-choice task using the left or right arrow key, responding as accurately and quickly as possible. Each image in the stimulus set was presented once. The experiment comprised ten blocks of 144 trials each, with self-paced breaks between blocks, and image order was randomized across trials. The session lasted approximately 30–40 minutes.

Analysis and Statistics. For each participant, we computed the individual FID-specific acceptance rate $P(real|FID, participant)$: the probability of a participant to perceive all images of a specific FID as real. For each image, we computed: (1) the realness $P(real|image)$ (the probability of an image to be perceived as real by all participants) and (2) the mean response time to an image (across participants). Real images were introduced to the statistical model with an image quality FID value of 0. The threshold for statistical significance was set at 0.05.

4. Results

Responses. High inter-person variability, well-documented for human perception judgments, is also evident in our data (Krippendorff's alpha = 0.301, N = 1440 images). Figure 2 displays traces of individual FID-specific acceptance rates with a general pattern: as FID scores increase, acceptance rates and inter-person variability decrease, indicating greater consensus for lower-quality images.

Across 20 participants, average human realness judgments follow image quality levels indicated by FID scores. Figure 3 displays the realness of an image, which generally decreases as FID scores increase.

We fitted a binomial logistic regression model to predict participants' realness classification (*real* vs. *fake*) from the image quality

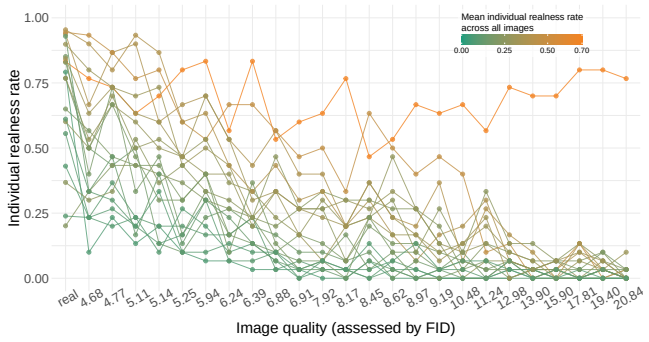

Figure 2: *Individual FID-specific acceptance rates (one trace per participant; each point: mean over 30 images). Color denotes the overall mean realness rate provided by each participant.*

Figure 3: *Perceived realness as a function of image quality. Dotted horizontal line: average realness across all real photographs. Black horizontal lines: medians; orange points: means. Yellow, red, and purple points correspond to images in Figure 4.*

(quantified by FID score). The model $R_i \sim q_i + (1|Participant)$ included image quality (q_i) as a fixed effect and participant as a random effect revealed a significant negative effect of FID on realness responses ($\beta = -0.32$, $SE = 0.004$, $z = -73.89$, $p < .001$), indicating that lower quality images were less likely to be classified as real, with notable heterogeneity across participants ($\sigma^2 = 0.98$). Figure 4 illustrates cases of strong agreement (yellow images) and divergence (red and purple images) between FID scores and human realness ratings. Qualitative inspection of FID–human disagreement cases (see Fig. 4) suggests three recurring failure modes: (1) localized high-frequency artifacts (e.g., skin texture irregularities) that minimally affect global feature statistics; (2) semantic accessories (e.g., glasses) whose perceptual salience is underestimated by Inception-based embeddings; (3) holistic plausibility mismatches, where globally coherent faces violate subtle facial norms detectable by humans, particularly in the periocular region.

FID's underestimation of perceptual salience of semantic accessories (e.g., glasses) exerted a bidirectional influence on realness ratings: poorly generated glasses reduced human ratings despite low FID scores (red images), whereas well-generated glasses increased perceived realness even when FID suggested lower image quality (purple images). Notably, in cases where FID and human

Figure 4: *Selected images with FID-human agreement on high quality (yellow), FID overestimating quality (red), and FID underestimating quality (purple). Corresponding data points are highlighted in Figure 3.*

judgments agreed on high image quality (yellow images), faces consistently lacked glasses, suggesting that eye-wear generation and skin texture may systematically drive FID–human correspondence.

Response Times. Response times provide a confidence-weighted measure of perceptual judgment, allowing us to distinguish ambiguous realism from confidently perceived artifacts. As shown in Figure 5, reaction times decreased with decreasing image quality (increasing FID) for correct fake classifications, but remained constant across FID levels for incorrect real classifications. A linear mixed model ($T_r \sim q_i * r_i + b_i + (1|Participant)$) revealed significant main effects of image quality (q_i), response (r_i), plus a significant interaction ($q_i \times r_i$: $\beta = -0.01497, p < .001$): fake responses were given with lower confidence for high-quality images (longer response time) but with substantially higher confidence for low quality images (shorter response time). Block number (b_i) was included as covariate to account for time-on-task effects. This suggests that binary *real/fake* accuracy alone underestimates perceptual uncertainty and that confidence-weighted evaluation reveals failure cases earlier than accuracy-based metrics.

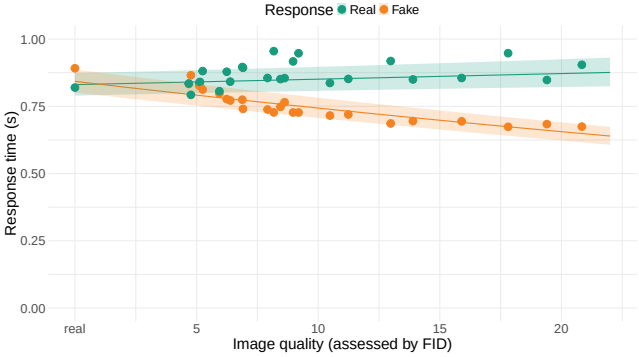

Figure 5: *Response times as a function of image quality.*

5. Discussion

We investigated the relation between FID scores and human judgments for GAN-generated face images.
Human judgments generally followed FID: As FID scores decreased, humans mainly perceived face images as more real, supporting that decreasing FID is related to increasing image quality. Further, with increasing FID (and decreasing image quality), correct responses (*fake*) were given with higher confidence, indexed by shorter response times. Interestingly, human realness ratings to actual photographs were, on average, still exceeded by those to high-quality synthetic faces. However, FID is not ideal: In cases where FID and human ratings did not agree, FID proved insensitive in identifying artifacts related to glasses (synthetic faces with the highest realness did not have glasses, while they were among those with lower FID scores) and unrealistic skin appearance. While these dichotomous ratings do not allow further use to improve the GAN model during the training process, continuous ratings might have this capacity. Brain activity recorded via electroencephalography (EEG) [CSB*24, BAS*18] might provide a promising source of continuous perceptual signals complementary to behavioral judgments. An EEG-based classifier, trained on a large dataset of such signals, could predict during GAN training whether generated images will be perceived as *real* or *fake* by humans. Our experiments deliberately focus on a single high-quality face generator, which allowed us to systematically vary image quality along a well-defined FID axis. While this limits the scope of the study, the identified failure modes are rooted in the structure of Inception-based feature embeddings and are therefore expected to generalize across GAN architectures trained on similar data. Extending this paradigm to diffusion models and other image domains is an important direction for future work. At the same time, we note that prolonged exposure can lead human raters to develop expertise in identifying generated images, which should be considered when designing perceptual evaluation protocols.

For the graphics community, this has direct implications for the evaluation of face synthesis and neural rendering systems: metrics optimized for global distribution matching may obscure perceptually critical local artifacts, particularly in high-fidelity regimes where realism judgments hinge on fine semantic detail.

Data Availability and Acknowledgement. The presented data are openly available at doi.org/10.12751/g-node.4gvwke. This project was funded by the Fraunhofer Society and the Max-Planck Society, project NeuroHum.

References

[BAS*18] BOSSE S., ACQUALAGNA L., SAMEK W., PORBADNIGK A. K., CURIO G., BLANKERTZ B., MULLER K.-R., WIEGAND T.: Assessing Perceived Image Quality Using Steady-State Visual Evoked Potentials and Spatio-Spectral Decomposition. *IEEE-TCSVT 28*, 8 (2018), 1694–1706. 4

[BMH*25] BARTHEL F., MORGENSTERN W., HINZER P., HILSMANN A., EISERT P.: Cgs-gan: 3d consistent gaussian splatting gans for high resolution human head synthesis. In *NeurIPS* (San Diego, US, Dec. 2025). 2

[Bor19] BORJI A.: Pros and cons of GAN evaluation measures. *CVIU 179* (2019), 41–65. 1, 2

[CSB*24] CHEN Y., STEPHANI T., BAGDASARIAN M. T., HILSMANN A., EISERT P., VILLRINGER A., BOSSE S., GAEBLER M., NIKULIN V. V.: Realness of face images can be decoded from non-linear modulation of EEG responses. *Sci Rep 14*, 1 (2024), 5683. 2, 4

[DL82] DOWSON D., LANDAU B.: The Fréchet distance between multivariate normal distributions. *J Multivar Anal 12*, 3 (1982), 450–455. 1

[Fré57] FRÉCHET M.: Sur la distance de deux lois de probabilité. In *Annales de l'ISUP* (1957), vol. 6, pp. 183–198. 1

[GBR*12] GRETTON A., BORGWARDT K. M., RASCH M. J., SCHÖLKOPF B., SMOLA A.: A kernel two-sample test. *J Mach Learn Res 13*, 25 (2012), 723–773. 2

[GPAM*14] GOODFELLOW I. J., POUGET-ABADIE J., MIRZA M., XU B., WARDE-FARLEY D., OZAIR S., COURVILLE A., BENGIO Y.: Generative adversarial nets. *NeurIPS 27* (2014). 1

[HRU*17] 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 *NeurIPS* (2017), vol. 30, pp. 6629–6640. 1, 2

[JRV*24] JAYASUMANA S., RAMALINGAM S., VEIT A., GLASNER D., CHAKRABARTI A., KUMAR S.: Rethinking FID: Towards a Better Evaluation Metric for Image Generation. In *IEEE-CVPR* (2024), pp. 9307–9315. 1, 2

[KLA19] KARRAS T., LAINE S., AILA T.: A style-based generator architecture for generative adversarial networks. In *IEEE-CVPR* (2019), pp. 4401–4410. 2

[KST*23] KAMMOUN A., SLAMA R., TABIA H., OUNI T., ABID M.: Generative Adversarial Networks for Face Generation. *ACM Comput Surv 55*, 5 (2023), 1–37. 1

[LKM*18] LUCIC M., KURACH K., MICHALSKI M., BOUSQUET O., GELLY S.: Are GANs created equal? a large-scale study. In *NeurIPS* (2018), vol. 31, pp. 698–707. 1

[LWLZ18] LIU S., WEI Y., LU J., ZHOU J.: An Improved Evaluation Framework for Generative Adversarial Networks, 2018. 1

[MBPW17] MOLLON J. D., BOSTEN J. M., PETERZELL D. H., WEBSTER M. A.: Individual differences in visual science. *Vis Res 141* (2017), 4–15. 1, 2

[SBK*13] SANDBERG K., BAHRAMI B., KANAI R., BARNES G. R., OVERGAARD M., REES G.: Early Visual Responses Predict Conscious Face Perception within and between Subjects during Binocular Rivalry. *J Cogn Neurosci 25*, 6 (2013), 969–985. 2

[SGZ*16] SALIMANS T., GOODFELLOW I., ZAREMBA W., CHEUNG V., RADFORD A., CHEN X., CHEN X.: Improved techniques for training gans. In *NeurIPS* (2016), vol. 29. 2

[SVI*16] SZEGEDY C., VANHOUCKE V., IOFFE S., SHLENS J., WOJNA Z.: Rethinking the inception architecture for computer vision. In *IEEE-CVPR* (2016), pp. 2818–2826. 1

[SZBK17] SCHINDLER S., ZELL E., BOTSCH M., KISSLER J.: Differential effects of face-realism and emotion on event-related brain potentials and their implications for the uncanny valley theory. *Sci Rep 7*, 1 (2017), 45003. 2

[TGB*17] TOLSTIKHIN I. O., GELLY S., BOUSQUET O., SIMON-GABRIEL C.-J., SCHÖLKOPF B.: AdaGAN: Boosting generative models. In *NeurIPS* (2017), vol. 30. 2

[ZGK*19] ZHOU S., GORDON M. L., KRISHNA R., NARCOMEY A., FEI-FEI L., BERNSTEIN M. S.: HYPE: A benchmark for human eye perceptual evaluation of generative models. In *NeurIPS* (2019). 2

[ZIE*18] ZHANG R., ISOLA P., EFROS A. A., SHECHTMAN E., WANG O.: The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR* (2018). 2