On the Robustness of Quality Measures for GANs

Motasem Alfarra[1], Juan C. Pérez[1], Anna Frühstück[1], Philip H. S. Torr[2], Peter Wonka[1], and Bernard Ghanem[1]

[1] King Abdullah University of Science and Tehchnology (KAUST), Saudi Arabia
[2] University of Oxford, United Kingdom
motasem.alfarra@kaust.edu.sa

Abstract. This work evaluates the robustness of quality measures of generative models such as Inception Score (IS) and Fréchet Inception Distance (FID). Analogous to the vulnerability of deep models against a variety of adversarial attacks, we show that such metrics can also be manipulated by additive pixel perturbations. Our experiments indicate that one can generate a distribution of images with very high scores but low perceptual quality. Conversely, one can optimize for small imperceptible perturbations that, when added to real world images, deteriorate their scores. We further extend our evaluation to generative models themselves, including the state of the art network StyleGANv2. We show the vulnerability of both the generative model and the FID against additive perturbations in the latent space. Finally, we show that the FID can be robustified by simply replacing the standard Inception with a robust Inception. We validate the effectiveness of the robustified metric through extensive experiments, showing it is more robust against manipulation.[3]

Keywords: Generative Adversarial Networks, Perceptual Quality, Adversarial Attacks, Network Robustness

1 Introduction

Deep Neural Networks (DNNs) are vulnerable to small imperceptible perturbations known as adversarial attacks. For example, while two inputs x and $(x + \delta)$ can be visually indistinguishable to humans, a classifier f can output two different predictions. To address this deficiency in DNNs, adversarial attacks [11,7] and defenses [20,27] have prominently emerged as active areas of research. Starting from image classification [28], researchers also assessed the robustness of DNNs for other tasks, such as segmentation [1], object detection [30], and point cloud classification [18]. While this lack of robustness questions the reliability of DNNs and hinders their deployment in the real world, DNNs are still widely used to evaluate performance in other computer vision tasks, such as that of generation.

Metrics in use for assessing generative models in general, and Generative Adversarial Networks (GANs) [10] in particular, are of utmost importance in the

[3] Code: https://github.com/R-FID-Robustness-of-Quality-Measures-for-GANs

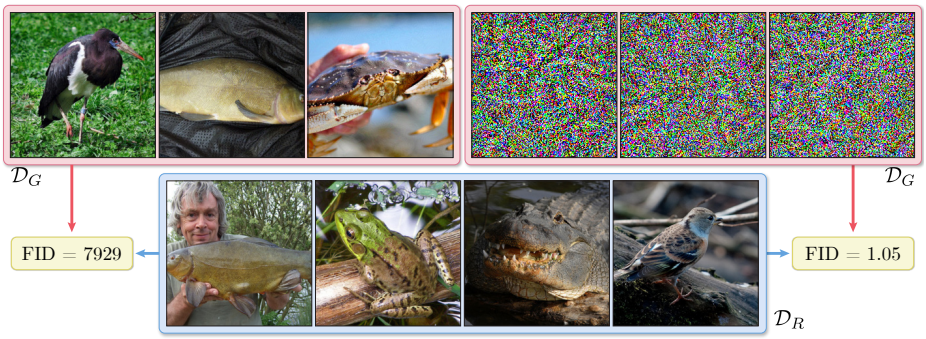

Fig. 1. Does the Fréchet Inception Distance (FID) accurately measure the distances between image distributions? We generate datasets that demonstrate the unreliability of FID in judging perceptual (dis)similarities between image distributions. The top left box shows a sample of a dataset constructed by introducing imperceptible noise to each ImageNet image. Despite the remarkable visual similarity between this dataset and ImageNet (bottom box), an extremely large FID (almost 8000) between these two datasets showcases FID's failure to capture perceptual similarities. On the other hand, a remarkably low FID (almost 1.0) between a dataset of random noise images (samples shown in the top right box) and ImageNet illustrates FID's failure to capture perceptual dissimilarities.

literature. This is because such metrics are widely used to establish the superiority of a generative model over others, hence guiding which GAN should be deployed in real world. Consequently, such metrics are expected to be not only useful in providing informative statistics about the distribution of generated images, but also reliable and robust. In this work, we investigate the robustness of metrics used to assess GANs. We first identify two interesting observations that are unique to this context. First, current GAN metrics are built on pretrained classification DNNs that are nominally trained (*i.e.* trained on clean images only). A popular DNN of choice is the Inception model [25], on which the Inception Score (IS) [22] and Fréchet Inception Distance (FID) [12] rely. Since nominally trained DNNs are generally vulnerable to adversarial attacks [7], it is expected that DNN-based metrics for GANs also inherit these vulnerabilities. Second, current adversarial attacks proposed in the literature are mainly designed at the instance level (*e.g.* fooling a DNN into misclassifying a particular instance), while GAN metrics are distribution-based. Therefore, attacking these distribution-based metrics requires extending attack formulations from the paradigm of instances to that of distributions.

In this paper, we analyze the robustness of GAN metrics and recommend solutions to improve their robustness. We first attempt to assess the robustness of the quality measures used to evaluate GANs. We check whether such metrics are actually measuring the quality of image distributions by testing their vulnerability against additive pixel perturbations. While these metrics aim at measuring perceptual quality, we find that they are extremely brittle against

imperceptible but carefully-crafted perturbations. We then assess the judgment of such metrics on the image distributions generated by StyleGANv2 [15] when its input is subjected to perturbations. While the output of GANs is generally well behaved, we still observe that such metrics provide inconsistent judgments where, for example, FID favors an image distribution with significant artifacts over more naturally-looking distributions. At last, we endeavor to reduce these metrics' vulnerability by incorporating robustly-trained models.

We summarize our contributions as follows:

- We are the first to provide an extensive experimental evaluation of the robustness of the Inception Score (IS) and the Fréchet Inception Distance (FID) against additive pixel perturbations. We propose two instance-based adversarial attacks that generate distributions of images that fool both IS and FID. For example, we show that perturbations δ with a small budget (*i.e.* $\|\delta\|_\infty \leq 0.01$) are sufficient to increase the FID between ImageNet [8] and a perturbed version of ImageNet to $\sim$ 7900, while also being able to generate a distribution of random noise images whose FID to ImageNet is 1.05. We illustrate both cases in Figure 1.
- We extend our evaluation to study the sensitivity of FID against perturbations in the latent space of state-of-the-art generative models. In this setup, we show the vulnerability of both StyleGANv2 and FID against perturbations in both its z- and w- spaces. We found that FID provides inconsistent evaluation of the distribution of generated images compared to their visual quality. Moreover, our attack in the latent space causes StyleGANv2 to generate images with significant artifacts, showcasing the vulnerability of StyleGANv2 to additive perturbations in the latent space.
- We propose to improve the reliability of FID by using adversarially-trained models in its computation. Specifically, we replace the traditional Inception model with its adversarially-trained counterpart to generate the embeddings on which the FID is computed. We show that our robust metric, dubbed R-FID, is more resistant against pixel perturbations than the regular FID.
- Finally, we study the properties of R-FID when evaluating different GANs. We show that R-FID is better than FID at distinguishing generated fake distributions from real ones. Moreover, R-FID provides more consistent evaluation under perturbations in the latent space of StyleGANv2.

2 Related Work

GANs and Automated Assessment. GANs [10] have shown remarkable generative capabilities, specially in the domain of images [14,15,4]. Since the advent of GANs, evaluating their generative capabilities has been challenging [10]. This challenge spurred research efforts into developing automated quantitative measures for GAN outputs. Metrics of particular importance for this purpose are the Inception Score (IS), introduced in [22], and the Fréchet Inception Distance (FID), introduced in [12]. Both metrics leverage the ImageNet-pretrained

Inception architecture [25] as a rough proxy for human perception. The IS evaluates the generated images by computing conditional class distributions with Inception and measuring (1) each distribution's entropy—related to Inception's certainty of the image content, and (2) the marginal's entropy—related to diversity across generated images. Noting the IS does not compare the generated distribution to the (real world) target distribution, Heusel *et al.* [12] proposed the FID. The FID compares the generated and target distributions by (1) assuming the Inception features follow a Gaussian distribution and (2) using each distribution's first two moments to compute the Fréchet distance. Further, the FID was shown to be more consistent with human judgement [24].

Both the original works and later research criticized these quantitative assessments. On one hand, IS is sensitive to weight values, noisy estimation when splitting data, distribution shift from ImageNet, susceptibility to adversarial examples, image resolution, difficulty in discriminating GAN performance, and vulnerability to overfitting [2,22,3,29]. On the other hand, FID has been criticized for its over-simplistic assumptions ("Gaussianity" and its associated two-moment description), difficulty in discriminating GAN performance, and its inability to detect overfitting [3,19,29]. Moreover, both IS and FID were shown to be biased to both the number of samples used and the model to be evaluated [6]. In this work, we provide extensive empirical evidence showing that both IS and FID are not robust against perturbations that modify image quality. Furthermore, we also propose a new *robust* FID metric that enjoys superior robustness.

Adversarial Robustness. While DNNs became the *de facto* standard for image recognition, researchers found that such DNNs respond unexpectedly to small changes in their input [26,11]. In particular, various works [5,20] observed a widespread vulnerability of DNN models against input perturbations that did not modify image semantics. This observation spurred a line of research on adversarial attacks, aiming to develop procedures for finding input perturbations that fool DNNs [7]. This line of work found that these vulnerabilities are pervasive, casting doubt on the nature of the impressive performances of DNNs. Further research showed that training DNNs to be robust against these attacks [20] facilitated the learning of perceptually-correlated features [13,9]. Interestingly, a later work [23] even showed that such learnt features could be harnessed for image synthesis tasks. In this work, we show (1) that DNN-based scores for GANs are vulnerable against adversarial attacks, and (2) how these scores can be "robustified" by replacing nominally trained DNNs with robustly trained ones.

3 Robustness of IS and FID

To compare the output of generative models, two popular metrics are used: the *Inception Score* (IS) and the *Fréchet Inception Distance* (FID). These metrics depend only on the statistics of the distribution of generated images in an ImageNet-pretrained Inception's embedding space, raising the question:

What do quality measures for generative models, such as IS and FID, tell us about image quality?

We investigate this question from the robustness perspective. In particular, we analyze the sensitivity of these metrics to carefully crafted perturbations. We start with preliminary background about both metrics.

3.1 Preliminaries

We consider the standard image generation setup where a generator $G : \mathbb{R}^{d_z} \to \mathbb{R}^{d_x}$ receives a latent code $z \in \mathbb{R}^{d_z}$ and outputs an image $x \in \mathbb{R}^{d_x}$. Upon training, G is evaluated based on the quality of the generated distribution of images $\mathcal{D}_G$ by computing either the IS [22] or the FID [12]. Both metrics leverage an ImageNet-pretrained [8] InceptionV3 [25]. Salimans *et al.* [22] proposed measuring the perceptual quality of the generated distribution $\mathcal{D}_G$ by computing the IS as:

$$\mathrm{IS}(\mathcal{D}_G) = \exp\left(\, \mathbb{E}_{x\sim\mathcal{D}_G} \left(\mathrm{KL}\left(p(y|x) \,||\, p(y)\right)\right)\right), \tag{1}$$

where $p(y|x)$ is the output probability distribution of the pretrained Inception model. While several works have argued about the effectiveness of the IS and its widely-used implementation [2], its main drawback is that it disregards the relation between the generated distribution, $\mathcal{D}_G$, and the real one, $\mathcal{D}_R$, used for training G [12]. Consequently, Heusel *et al.* proposed the popular FID, which involves the statistics of the real distribution. In particular, FID assumes that the Inception features of an image distribution $\mathcal{D}$ follow a Gaussian distribution with mean $\mu_{\mathcal{D}}$ and covariance $\Sigma_{\mathcal{D}}$, and it measures the squared Wasserstein distance between the two Gaussian distributions of real and generated images. Hence, $\mathrm{FID}(\mathcal{D}_R, \mathcal{D}_G)$, or FID for short, can be calculated as:

$$\mathrm{FID} = \|\mu_R - \mu_G\|^2 + \mathrm{Tr}\left(\Sigma_R + \Sigma_G - 2(\Sigma_R\Sigma_G)^{1/2}\right), \tag{2}$$

where $\cdot_R, \cdot_G$ are the statistics of the real and generated image distributions, respectively, and $\mathrm{Tr}(\cdot)$ is the trace operator. Note that the statistics of both distributions are empirically estimated from their corresponding image samples. In principle, FID measures how close (realistic) the generated distribution $\mathcal{D}_G$ is to $\mathcal{D}_R$. We remark that the FID is the *de facto* metric for evaluating image generation-related tasks. Therefore, our study focuses mostly on FID.

We note here that both the IS and the FID are oblivious to G's training process and can be computed to compare two arbitrary sets of images $\mathcal{D}_R$ and $\mathcal{D}_G$. In generative modeling, this is typically a set of real images (photographs) and a set of generated images. However, it is also possible to compare two sets of photographs, two sets of generated images, manipulated photographs with real photographs, *etc.* This flexibility allows us to study these metrics in a broader context next, where no generative model is involved.

3.2 Robustness under Pixel Perturbations

We first address the question presented earlier in Section 3 by analyzing the sensitivity of IS and FID to additive pixel perturbations. In particular, we assume

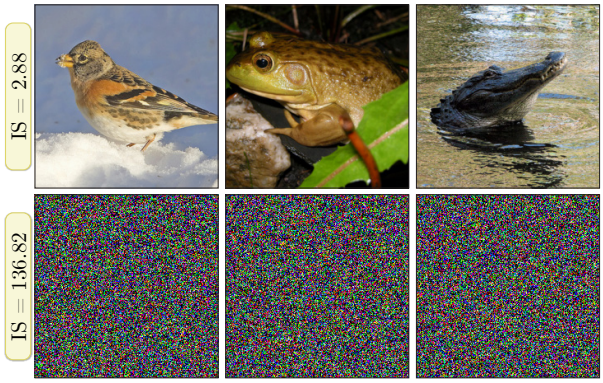

Fig. 2. Sensitivity of Inception Score (IS) against pixel perturbations. *First row:* real-looking images (sampled from $\mathcal{D}_G = \mathcal{D}_R + \delta$) with a low IS (below 3). *Second row:* random noise images with a high IS (over 135).

$\mathcal{D}_R$ to be either CIFAR10 [17] or ImageNet [8] and ask: **(i)** can we generate a distribution of imperceptible additive perturbations δ that deteriorates the scores for $\mathcal{D}_G = \mathcal{D}_R + \delta$? Or, alternatively, **(ii)** can we generate a distribution of low visual quality images, *i.e.* noise images, that attain good quality scores? If the answer is yes to both questions, then FID and IS have limited capacity for providing information about image quality in the worst case.

Good Images - Bad Scores We aim at constructing a distribution of real-looking images with *bad* quality measures, *i.e.* low IS or high FID. While both metrics are distribution-based, we design instance-wise proxy optimization problems to achieve our goal.

Minimizing IS. Based on Eq. (1), one could minimize the IS by having both the posterior $p(y|x)$ and the prior $p(y)$ be the same distribution. Assuming that $p(y)$ is a uniform distribution, we minimize the IS by maximizing the entropy of $p(y|x)$. Therefore, we can optimize a perturbation δ^* for each real image $x_r \sim \mathcal{D}_R$ by solving the following problem:

$$\begin{aligned} \delta^* &= \underset{\|\delta\|_\infty \leq \epsilon}{\arg\max}\ \mathcal{L}_{\text{ce}}\left(p(y|x_r+\delta), \hat{y}\right), \\ \text{s.t. } \hat{y} &= \underset{i}{\arg\max}\ p^i(y|x_r+\delta), \end{aligned} \tag{3}$$

where $\mathcal{L}_{\text{ce}}$ is the Cross Entropy loss. We solve the problem in Eq. (3) with 100 steps of Projected Gradient Descent (PGD) and zero initialization. We then compile the distribution $\mathcal{D}_G$, where each image $x_g = x_r + \delta^*$ is a perturbed version of the real dataset $\mathcal{D}_R$. Note that our objective aims to minimize the network's confidence in predicting all labels for each x_g. In doing so, both $p(y|x_g)$ and $p(y)$ tend to converge to a uniform distribution, thus, minimizing the KL divergence between them and effectively lowering the IS. Note how ϵ controls the

Table 1. Robustness of IS and FID against pixel perturbations. We assess the robustness of IS and FID against perturbations with a limited budget ϵ on CIFAR10 and ImageNet. In the last row, we report the IS and FID of images with carefully-designed random noise having a resolution similar to CIFAR10 and ImageNet.

ϵ	CIFAR10		ImageNet	
	IS	FID	IS	FID
0.00	11.54	0.00	250.74	0.00
5×10^{-3}	2.62	142.45	3.08	3013.33
0.01	2.50	473.19	2.88	7929.01
random noise	94.87	9.94	136.82	1.05

allowed perturbation amount for each image x_r. Therefore, for small ϵ values, samples from $\mathcal{D}_G$ and $\mathcal{D}_R$ are perceptually indistinguishable.

Maximizing FID. Next, we extend our attack setup to the more challenging FID. Given an image x, we define $f(x) : \mathbb{R}^{d_x} \to \mathbb{R}^{d_e}$ to be the output embedding of an Inception model. We aim to maximize the FID by generating a perturbation δ that pushes the embedding of a real image away from its original position. In particular, for each $x_r \sim \mathcal{D}_R$, we aim to construct $x_g = x_r + \delta^*$ where:

$$\delta^* = \underset{\|\delta\|_\infty \leq \epsilon}{\arg\max} \; \|f(x_r) - f(x_r + \delta)\|_2 \,. \tag{4}$$

In our experiments, we solve the optimization problem in Eq. (4) with 100 PGD steps and a randomly initialized δ [20]. Maximizing this objective indirectly maximizes FID's first term (Eq. (2)), while resulting in a distribution of images $\mathcal{D}_G$ that is visually indistinguishable from the real $\mathcal{D}_R$ for small ϵ values.

Experiments. We report our results in Table 1. Our simple yet effective procedure illustrates how both metrics are very susceptible to attacks. In particular, solving the problem in Eq. (3) yields a distribution of noise that significantly decreases the IS from 11.5 to 2.5 in CIFAR10 and from 250.7 to 2.9 in ImageNet. We show a sample from $\mathcal{D}_G$ in Figure 2, first row. Similarly, our optimization problem in Eq. (4) can create imperceptible perturbations that maximize the FID to ≈7900 between ImageNet and its perturbed version (examples shown in Figure 1).

Bad Images - Good Scores While the previous experiments illustrate the vulnerability of both the IS and FID against small perturbations (*i.e.* good images with bad scores), here we evaluate if the converse is also possible, *i.e.* bad images with good scores. In particular, we aim to construct a distribution of noise images (*e.g.* second row of Figure 2) that enjoys good scores (high IS or low FID).

Maximizing IS. The IS has two terms: Inception's confidence on classifying a generated image, *i.e.* $p(y|x_g)$, and the diversity of the generated distribution of predicted labels, *i.e.* $p(y)$. One can maximize the IS by generating a distribution $\mathcal{D}_G$ such that: **(i)** each $x_g \sim \mathcal{D}_G$ is predicted with high confidence, and **(ii)** the distribution of predicted labels is uniform across Inception's output $\mathcal{Y}$. To that

end, we propose the following procedure for constructing such $\mathcal{D}_G$. For each x_g, we sample a label $\hat{y} \sim \mathcal{Y}$ uniformly at random and solve the problem:

$$x_g = \arg\min_x \ \mathcal{L}_{ce}(p(y|x), \hat{y}). \tag{5}$$

In our experiments, we solve the problem in Eq. (5) with 100 gradient descent steps and random initialization for x.
Minimizing FID. Here, we analyze the robustness of FID against such a threat model. We follow a similar strategy to the objective in Eq. (4). For each image $x_r \sim \mathcal{D}_R$, we intend to construct x_g such that:

$$x_g = \arg\min_x \ \|f(x) - f(x_r)\|_2 \tag{6}$$

with a randomly initialized x. In our experiments, we solve Eq. (6) with 100 gradient descent steps. As such, each x_g will have a similar Inception representation to a real-world image, *i.e.* $f(x_g) \approx f(x_r)$, while being random noise.

Experiments. We report our results in the last row of Table 1. Both the objectives in Eqs. (5) and (6) are able to fool the IS and FID, respectively. In particular, we are able to generate distributions of noise images with resolutions 32×32 and 224×224 (*i.e.* CIFAR10 and ImageNet resolutions) but with IS of 94 and 136, respectively. We show a few qualitative samples in the second row of Figure 2. Furthermore, we generate noise images that have embedding representations very similar to those of CIFAR10 and ImageNet images. This lowers the FID of both datasets to 9.94 and 1.05, respectively (examples are shown in Figure 1).

3.3 Robustness under Latent Perturbations

In the previous section, we established the vulnerability of both the IS and FID against pixel perturbations. Next, we investigate the vulnerability against perturbations in a GAN's latent space. Designing such an attack is more challenging in this case, since images can only be manipulated indirectly, and so there are fewer degrees of freedom for manipulating an image. To that end, we choose G to be the state of the art generator StyleGANv2 [14] trained on the standard FFHQ dataset [14]. We limit the investigation to the FID metric, as IS is not commonly used in the context of unconditional generators, such as StyleGAN. Note that we always generate 70k samples from G to compute the FID.

Recall that our generator G accepts a random latent vector $z \sim \mathcal{N}(0, \mathrm{I})$[4] and maps it to the more expressive latent space w, which is then fed to the remaining layers of G. It is worthwhile to mention that "truncating" the latent w with a pre-computed $\bar{w}$[5] and constant $\alpha \in \mathbb{R}$ (*i.e.* replacing w with $\alpha w + (1-\alpha)\bar{w}$) controls both the quality and diversity of the generated images [14].

[4] The appendix presents results showing that sampling z from different distributions still yields good looking StyleGANv2-generated images.
[5] $\bar{w}$ is referred to as the mean of the w-space. It is computed by sampling several latents z and averaging their representations in the w-space.

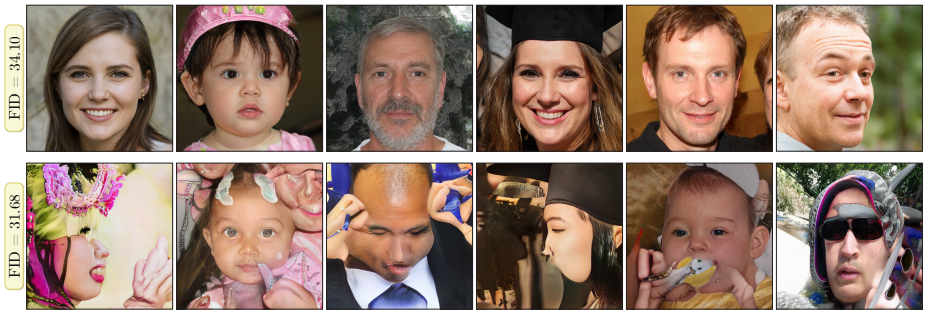

Fig. 3. Effect of attacking truncated StyleGANv2's latent space on the Fréchet Inception Distance (FID). We conduct attacks on the latent space of StyleGANv2 and record the effect on the FID. We display the resulting samples of these attacks for two truncation values, $\alpha = 0.7$ (top row) and $\alpha = 1.0$ (bottom row). Despite the stark differences in realism between the images in the top and bottom rows—*i.e.* the top row's remarkable quality and the bottom row's artifacts—the FID to FFHQ reverses this ranking, wherein the bottom row is judged as *farther* away from FFHQ than the top row.

Effect of Truncation on FID. We first assess the effect of the truncation level α on both image quality and FID. We set $\alpha \in [0.7, 1.0, 1.3]$ and find FIDs to be $[21.81, 2.65, 9.31]$, respectively. Based on our results, we assert the following observation: while the visual quality of generated images at higher truncation levels, *e.g.* $\alpha = 0.7$, is better and has fewer artifacts than the other α values, the FID does not reflect this fact, showing lower (better) values for $\alpha \in \{1.0, 1.3\}$. We elaborate on this observation with qualitative experiments in the appendix.

FID-Guided Sampling. Next, we extend the optimization problem in Eq. (4) from image to latent perturbations. In particular, we aim at constructing a perturbation δ_z^* for each sampled latent z by solving:

$$\delta_z^* = \arg\max_{\delta} \; \|f(G(z+\delta)) - f(x_r)\|_2 \,. \tag{7}$$

Thus, δ_z^* perturbs z such that G produces an image whose embedding differs from that of real image x_r. We solve the problem in Eq. (7) for $\alpha \in \{0.7, 1.0\}$.

Experiments. We visualize our results in Figure 3 accompanied with their corresponding FID values (first and second rows correspond to $\alpha = 0.7$ and 1.0, respectively). While our attack in the latent space is indeed able to significantly increase the FID (from 2.65 to 31.68 for $\alpha = 1.0$ and 21.33 to 34.10 for $\alpha = 0.7$), we inspect the results and draw the following conclusions. **(i)** FID provides inconsistent evaluation of the generated distribution of images. For example, while both rows in Figure 3 have comparable FID values, the visual quality is significantly different. This provides practical evidence of this metric's unreliability in measuring the performance of generative models. **(ii)** Adding crafted perturbations to the input of a state of the art GAN deteriorates the visual quality of its

output space (second row in Figure 3). This means that GANs are also vulnerable to adversarial attacks. This is confirmed in the literature for other generative models such as GLOW [16,21]. Moreover, we can formulate a problem similar to Eq. (7) but with the goal of perturbing the w-space instead of the z-space. We leave results of solving this formulation for different α values to the appendix.

Section Summary. In this section, we presented an extensive experimental evaluation investigating if the quality measures (IS and FID) of generative models actually measure the perceptual quality of the output distributions. We found that such metrics are extremely vulnerable to pixel perturbations. We were able to construct images with very good scores but no visual content (Section 3.2), as well as images with realistic visual content but very bad scores (Section 3.2). We further studied the sensitivity of FID against perturbations in the latent space of StyleGANv2 (Section 3.3), allowing us to establish the inconsistency of FID under this setup as well. Therefore, we argue that such metrics, while measuring useful properties of the generated distribution, lead to questionable assessments of the visual quality of the generated images.

4 R-FID: Robustifying the FID

After establishing the vulnerability of IS and FID to perturbations, we analyze the cause of such behavior and propose a solution. We note that, while different metrics have different formulations, they rely on a pretrained Inception model that could potentially be a leading cause of such vulnerability. This observation suggests the following question:

Can we robustify the FID by replacing its Inception component with a robustly trained counterpart?

We first give a brief overview of adversarial training.

4.1 Leveraging Adversarially Trained Models

Adversarial training is arguably the *de facto* procedure for training robust models against adversarial attacks. Given input-label pairs (x, y) sampled from a training set $\mathcal{D}_{tr}$, ℓ_2-adversarial training solves the following min-max problem:

$$\min_{\theta} \; \mathbb{E}_{(x,y)\sim\mathcal{D}_{tr}} \left[\max_{\|\delta\|_2 \leq \kappa} \mathcal{L}\left(x+\delta, y; \theta\right) \right] \tag{8}$$

for a given loss function $\mathcal{L}$ to train a robust network with parameters θ. We note that κ controls the robustness-accuracy trade-off: models trained with larger κ tend to have higher robust accuracy (accuracy under adversarial attacks) and lower clean accuracy (accuracy on clean images). Since robust models are expected to resist pixel perturbations, we expect such models to inherit robustness characteristics against the attacks constructed in Section 3.2. Moreover, earlier

Fig. 4. Attacking R-FID with pixel perturbations. We attack two variants of R-FID ($\kappa = 64$ and $\kappa = 128$) and visualize samples from the resulting datasets. Attempting to fool these R-FIDs at the pixel level yields perturbations that correlate with semantic patterns, in contrast to those obtained when attempting to fool the standard FID (as shown in Figure 1).

Table 2. R-FID against attacks in the pixel space. We study the robustness of R-FID against the adversarial attacks in Eq. 4.

ϵ	CIFAR10		ImageNet	
	$\kappa = 64$	$\kappa = 128$	$\kappa = 64$	$\kappa = 128$
0.01	1.5	0.3	21.0	4.5
0.02	20.7	7.8	293.8	92.1
0.03	46.4	19.7	657.9	264.6

works showed that robustly-trained models tend to learn more semantically-aligned and invertible features [13]. Therefore, we hypothesize that replacing the pretrained Inception model with its robustly trained counterpart could increase FID's sensitivity to the visual quality of the generated distribution (*i.e.* robust against attacks in Section 3.3).

To that end, we propose the following modification to the FID computation. We replace the pretrained Inception model with a robustly trained version on ImageNet following Eq. (8) with $\kappa \in \{64, 128\}$. The training details are left to the appendix. We refer to this alternative as R-FID, and analyze its robustness against perturbations next.

4.2 R-FID against Pixel Perturbations

We first test the sensitivity of R-FID against additive pixel perturbations. For that purpose, we replace the Inception with a robust Inception, and repeat the experiments from Section 3.2 to construct real images with bad scores. We con-

Table 3. Truncation's effect on R-FID. We study how truncation affects R-FID against FFHQ (first two rows), and across different truncation levels (last two rows).

$(\mathcal{D}_G(\alpha), \mathcal{D}_R)$	0.7	0.9	1.0
$\kappa = 64$	98.3	90.0	88.1
$\kappa = 128$	119.9	113.7	113.8
$(\mathcal{D}_G(\alpha_i), \mathcal{D}_G(\alpha_j))$	(0.7, 1.0)	(0.7, 0.9)	(0.9, 1.0)
$\kappa = 64$	10.5	4.9	0.48
$\kappa = 128$	9.9	4.6	0.46

duct experiments on CIFAR10 and ImageNet with $\epsilon \in \{0.01, 0.02, 0.03\}$ for the optimization problem in Eq. (4), and we report the results in Table 2. We observe that the use of a robustly-trained Inception significantly improves robustness against pixel perturbations. Our robustness improvement for the same value of $\epsilon = 0.01$ is of 3 orders of magnitude (an FID of 4 for $\kappa = 128$ compared to 7900 reported in Table 1). While both models consistently provide a notable increase in robustness against pixel perturbations, we find that the model most robust to adversarial attacks (*i.e.* $\kappa = 128$) is also the most robust to FID attacks. It is worthwhile to mention that this kind of robustness is expected since our models are trained not to alter their prediction under additive input perturbations. Hence, their feature space should enjoy robustness properties, as measured by our experiments. In Figure 4 we visualize a sample from the adversarial distribution $\mathcal{D}_G$ (with $\epsilon = 0.08$) when $\mathcal{D}_R$ is ImageNet. We observe that our adversaries while aiming only at pushing the feature representation of samples of $\mathcal{D}_G$ away from those of $\mathcal{D}_R$, are also more correlated with human perception. This finding aligns with previous observations in the literature, which find robustly-trained models have a more interpretable (more semantically meaningful) feature space [13,9]. We leave the evaluation under larger values of ϵ, along with experiments on unbounded perturbations, to the appendix.

4.3 R-FID under Latent Perturbations

In Section 4.2, we tested R-FID's robustness against pixel-level perturbations. Next, we study R-FID for evaluating generative models. For this, we follow the setup in Section 3.3 using an FFHQ-trained StyleGANv2 as generator G.

Effect of Truncation on R-FID. Here, we analyze the R-FID when the generator is using different truncation levels. In particular, we choose $\alpha \in \{0.7, 0.9, 1.0\}$ and report results in Table 3. We observe that the robust Inception model clearly distinguishes the distribution generated by StyleGANv2 from the FFHQ dataset, regardless of the truncation α. In this case, we obtain an R-FID of 113.8, substantially larger than the 2.6 obtained when the nominally-trained Inception model is used. This result demonstrates that, while the visual quality of StyleGANv2's output is impressive, the generated image distribution is far from the FFHQ

Fig. 5. Robustness of R-FID against perturbations in StyleGANv2 latent space. We conduct attacks on two variants of R-FID ($\kappa = 64$ on the left, and $\kappa = 128$ on the right) and two truncation values ($\alpha = 0.7$ on the top, and $\alpha = 1.0$ on the bottom) by perturbing the latent space. We also visualize samples from the generated distributions. For the pairs $(\kappa, \alpha) \in \{(64, 0.7), (64, 1.0), (128, 0.7), (128, 1.0)\}$, we find corresponding R-FID values of {128.1, 157.8, 126.6, 162,8}. In contrast to the minimal changes required to fool the standard FID (Fig. 3), fooling the R-FID leads to a dramatic degradation in visual quality of the generated images.

distribution. We further evaluate if the R-FID is generally large between any two distributions by measuring the R-FID between two distributions of images generated at two truncation levels (α_i, α_j). Table 3 reports these results. We observe that **(i)** the R-FID between a distribution and itself is ≈ 0, *e.g.* R-FID $= 10^{-3}$ at (1.0, 1.0). Please refer to the appendix for details. **(ii)** The R-FID gradually increases as the image distributions differ, *e.g.* R-FID at (0.9, 1.0) $<$ (0.7, 1.0). This observation validates that the large R-FID values found between FFHQ and various truncation levels are a result of the large separation in the embedding space that robust models induce between real and generated images.

R-FID Guided Sampling. Next, we assess the robustness of the R-FID against perturbations in the latent space of the generator G. For this purpose, we conduct the attack proposed in Eq. (7) with f now being the robustly-trained Inception. We report results and visualize few samples in Figure 5. We make the following observations. **(i)** While the R-FID indeed increases after the attack, the relative increment is far less than that of the non-robust FID. For example, R-FID increases by 44% at $\kappa = 64$ and $\alpha = 0.7$ compared to an FID increase of 1000% under the same setup. **(ii)** The increase in R-FID is associated with a significantly larger amount of artifacts introduced by the GAN in the generated images. This result further evidences the vulnerability of the generative model. However, it also highlights the changes in the image distribution that are required to increase the R-FID. We leave the w- space formulation for the attack on the R-FID, along with its experiments, to the appendix.

Section Summary. In this section, we robustified the popular FID by replacing the pretrained Inception model with a robustly-trained version. We found this

Table 4. Sensitivity of R-FID against noise and blurring. We measure R-FID ($\kappa = 128$) between ImageNet and a transformed version of it under Gaussian noise and blurring. As σ increases, the image quality decreases and R-FID increases.

σ_N/σ_B	$0.1/1.0$	$0.2/2.0$	$0.3/3.0$	$0.4/4.0$
Gaussian (N)oise	16.65	61.33	128.8	198.3
Gaussian (B)lur	15.54	54.07	78.67	89.11

replacement results in a more robust metric (R-FID) against perturbations in both the pixel (Section 4.2) and latent (Section 4.3) spaces. Moreover, we found that pixel-based attacks yield much more perceptually-correlated perturbations when compared to the attacks that used the standard FID (Figure 2). Finally, we observed that changing R-FID values requires a more significant and notable distribution shift in the generated images (Figure 5).

4.4 R-FID against Quality Degradation

At last, we analyze the effect of transformations that degrade image quality on R-FID. In particular, we apply Gaussian noise and Gaussian blurring on ImageNet and report the R-FID ($\kappa = 128$) between ImageNet and the degraded version in Table 4. Results show that as the quality of the images degrades (*i.e.* as σ increases), the R-FID steadily increases. Thus, we find that R-FID is able to distinguish a distribution of images from its degraded version.

5 Discussion, Limitations, and Conclusions

In this work, we demonstrate several failure modes of popular GAN metrics, specifically IS and FID. We also propose a robust counterpart of FID (R-FID), which mitigates some of the robustness problems and yields significantly more robust behavior under the same threat models.

Measuring the visual quality for image distributions has two components: (1) the statistical measurement (*e.g.* Wasserstein distance) and (2) feature extraction using a pretrained model (*e.g.* InceptionV3). A limitation of our work is that we only focus on the second part (the pretrained model). As an interesting avenue for future work, we suggest a similar effort to assess the reliability of the statistical measurement as well, *i.e.* analyzing and finding better and more robust alternatives to the Wasserstein distance.

Current metrics mainly focus on comparing the distribution of features. In these cases, visual quality is only hoped to be a side effect and not directly optimized for nor tested by these metrics. Developing a metric that directly assesses visual quality remains an open problem that is not tackled by our work but is recommended for future work.

Acknowledgments. This work was supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. OSR-CRG2019-4033.

References

1. Arnab, A., Miksik, O., Torr, P.H.: On the robustness of semantic segmentation models to adversarial attacks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 888–897 (2018)
2. Barratt, S., Sharma, R.: A note on the inception score (2018)
3. Borji, A.: Pros and cons of gan evaluation measures. Computer Vision and Image Understanding **179**, 41–65 (2019)
4. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: International Conference on Learning Representations (ICLR) (2019)
5. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP) (2017)
6. Chong, M.J., Forsyth, D.: Effectively unbiased fid and inception score and where to find them. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6070–6079 (2020)
7. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International Conference on Machine Learning (ICML) (2020)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F.: Imagenet: A large-scale hierarchical image database. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
9. Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Tran, B., Madry, A.: Adversarial robustness as a prior for learned representations (2020), https://openreview.net/forum?id=rygvFyrKwH
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (2014)
11. Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (ICLR) (2015)
12. 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: Advances in Neural Information Processing Systems (NeurIPS) (2017)
13. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial examples are not bugs, they are features. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
14. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
15. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
16. Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018), https://proceedings.neurips.cc/paper/2018/file/d139db6a236200b21cc7f752979132d0-Paper.pdf
17. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. In: University of Toronto, Canada (2009)

18. Liu, H., Jia, J., Gong, N.Z.: Pointguard: Provably robust 3d point cloud classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6186–6195 (2021)
19. Lucic, M., Kurach, K., Michalski, M., Gelly, S., Bousquet, O.: Are gans created equal? a large-scale study. Advances in Neural Information Processing Systems (NeurIPS) **31** (2018)
20. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (ICLR) (2018)
21. Pope, P., Balaji, Y., Feizi, S.: Adversarial robustness of flow-based generative models. In: Chiappa, S., Calandra, R. (eds.) Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 108, pp. 3795–3805. PMLR (26–28 Aug 2020)
22. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., Chen, X.: Improved techniques for training gans. In: Advances in Neural Information Processing Systems (NeurIPS) (2016)
23. Santurkar, S., Ilyas, A., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Image synthesis with a single (robust) classifier. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
24. Shmelkov, K., Schmid, C., Alahari, K.: How good is my gan? In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 213–229 (2018)
25. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (2016)
26. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: International Conference on Learning Representations (ICLR) (2014)
27. Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., Gu, Q.: Improving adversarial robustness requires revisiting misclassified examples. In: International Conference on Learning Representations (ICLR) (2019)
28. Wu, D., Xia, S.T., Wang, Y.: Adversarial weight perturbation helps robust generalization. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
29. Xu, Q., Huang, G., Yuan, Y., Guo, C., Sun, Y., Wu, F., Weinberger, K.: An empirical study on evaluation metrics of generative adversarial networks. arXiv preprint arXiv:1806.07755 (2018)
30. Zhao, Y., Zhu, H., Liang, R., Shen, Q., Zhang, S., Chen, K.: Seeing isn't believing: Towards more robust adversarial attack against real world object detectors. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. pp. 1989–2004 (2019)