

Pros and Cons of GAN Evaluation Measures

Ali Borji

aliborji@gmail.com

Abstract

Generative models, in particular generative adversarial networks (GANs), have gained significant attention in recent years. A number of GAN variants have been proposed and have been utilized in many applications. Despite large strides in terms of theoretical progress, evaluating and comparing GANs remains a daunting task. While several measures have been introduced, as of yet, there is no consensus as to which measure best captures strengths and limitations of models and should be used for fair model comparison. As in other areas of computer vision and machine learning, it is critical to settle on one or few good measures to steer the progress in this field. In this paper, I review and critically discuss more than 24 quantitative and 5 qualitative measures for evaluating generative models with a particular emphasis on GAN-derived models. I also provide a set of 7 desiderata followed by an evaluation of whether a given measure or a family of measures is compatible with them.

Keywords: Generative Adversarial Nets, Generative Models, Evaluation, Deep Learning, Neural Networks

1. Introduction

Generative models are a fundamental component of a variety of important machine learning and computer vision algorithms. They are increasingly used to estimate the underlying statistical structure of high dimensional signals and artificially generate various kinds of data including high-quality images, videos, and audio. They can be utilized for purposes such as representation learning and semi-supervised learning [1, 2, 3], domain adaptation [4, 5], text to image synthesis [6], compression [7], super resolution [8], inpainting [9, 10], saliency prediction [11], image enhancement [12], style transfer and texture synthesis [13, 14], image-to-image translation [15, 16], and video generation and prediction [17]. A recent class of generative models known as Generative Adversarial Networks (GANs) by Goodfellow *et al.* [18] has attracted much attention. A sizable volume of follow-up papers have been published since introduction of GANs in 2014. There has been substantial progress in terms of theory and applications and a large number of GAN variants have been introduced. However, relatively less effort has been spent in evaluating GANs

and grounded ways to quantitatively and qualitatively assess them are still missing.

Generative models can be classified into two broad categories of *explicit* and *implicit* approaches. The former class assumes access to the model likelihood function, whereas the latter uses a sampling mechanism to generate data. Examples of explicit models are variational auto-encoders (VAEs) [19, 20] and PixelCNN [21]. Examples of implicit generative models are GANs. Explicit models are typically trained by maximizing the likelihood or its lower bound. GANs aim to approximate a data distribution P , using a parameterized model distribution Q . They achieve this by jointly optimizing two adversarial networks: a generator and a discriminator. The generator G is trained to synthesize from a noise vector an image that is close to the true data distribution. The discriminator D is optimized to accurately distinguish between the synthesized images coming from the generator and the real images from the data distribution. GANs have shown a dramatic ability to generate realistic high resolution images.

Several evaluation measures have surfaced with the emergence of new models. Some of them attempt to quantitatively evaluate models while some others emphasize on qualitative ways such as user studies or analyzing internals of models. Both of these approaches have strengths and limitations. For example, one may think that fooling a person in distinguishing generated images from real ones can be the ultimate test. Such a measure, however, may favor models that concentrate on limited sections of the data (*i.e.* overfitting or memorizing; low diversity; mode dropping). Quantitative measures, while being less subjective, may not directly correspond to how humans perceive and judge generated images. These, along with other issues such as the variety of probability criteria and the lack of a perceptually meaningful image similarity measures, have made evaluating generative models notoriously difficult [22]. In spite of no agreement regarding the best GAN evolution measure, few works have already started to benchmark GANs (*e.g.* [23, 24, 25]). While such studies are indeed helpful, further research is needed to understand GAN evaluation measures and assess their strengths and limitations (*e.g.* [22, 26, 27, 28, 29, 30]).

My main goal in this paper is to critically review available GAN measures and help the researchers objectively assess them. At the end, I will offer some suggestions for designing more efficient measures for fair GAN evaluation and comparison.

2. GAN Evaluation Measures

I will enumerate the GAN evaluation measures while discussing their pros and cons. They will be organized in two categories: *quantitative* and *qualitative*. Notice that some of these measures (*e.g.* Wasserstein distance, reconstruction error or SSIM) can also be used for model optimization during training. In the next subsection, I will first provide a set of desired properties for GAN measures (*a.k.a* meta measures or desiderata) followed by an evaluation of whether a given measure or a family of measures is compatible with them.

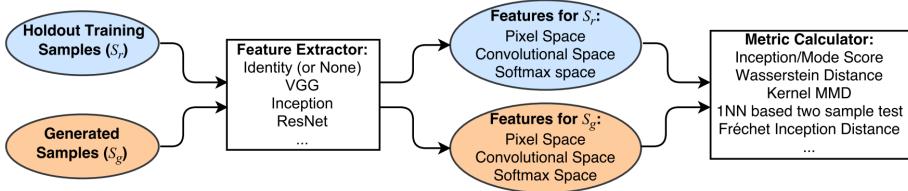


Figure 1: A schematic layout of the typical approach for sample based GAN evaluation. S_r and S_g represent real and generated samples, respectively. Figure from [26].

Table 1 shows the list of measures. Majority of the measures return a single value while few (GAM [31] and NRDS [32]) perform relative comparison. The rationale behind the latter is that if it is difficult to obtain the perfect measure, at least we can evaluate which model generates better images than others.

2.1. Desiderata

Before delving into the explanation of evaluation measures, first I list a number of desired properties that an efficient GAN evaluation measure should fulfill. These properties can serve as meta measures to evaluate and compare the GAN evaluation measures. Here, I emphasize on the qualitative aspects of these measures. As will be discussed in Section 3, some recent works have attempted to compare the meta measures quantitatively (*e.g.* computational complexity of a measure). An efficient GAN evaluation measure should:

1. favor models that generate high fidelity samples (*i.e.* ability to distinguish generated samples from real ones; discriminability),
2. favor models that generate diverse samples (and thus is sensitive to overfitting, mode collapse and mode drop, and can undermine trivial models such as the memory GAN),
3. favor models with disentangled latent spaces as well as space continuity (*a.k.a* controllable sampling),
4. have well-defined bounds (lower, upper, and chance),
5. be sensitive to image distortions and transformations. GANs are often applied to image datasets where certain transformations to the input do not change semantic meanings. Thus, an ideal measure should be invariant to such transformations. For instance, score of a generator trained on CelebA face dataset should not change much if its generated faces are shifted by a few pixels or rotated by a small angle.
6. agree with human perceptual judgments and human rankings of models, and
7. have low sample and computational complexity.

In what follows, GAN measures will be discussed and assessed with respect to the above desiderata, and a summary will be presented eventually in Section 3. See Table 2.

2.2. Quantitative Measures

A schematic layout for sample based GAN evaluation measures is shown in Fig. 1. Some measures discussed in the following are “model agnostic” in that the generator is used as a black box to sample images and they do not require a density estimation from the model. On the contrary, some other measures such as average log-likelihood demand estimating a probability distribution from samples.

1. **Average Log-likelihood.** Kernel density estimation (KDE or Parzen window estimation) is a well-established method for estimating the density function of a distribution from samples¹. For a probability kernel K (most often an isotropic Gaussian) and i.i.d samples $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, a density function at \mathbf{x} is defined as $p(\mathbf{x}) \approx \frac{1}{z} \sum_{i=1}^n K(\mathbf{x} - \mathbf{x}_i)$, where z is a normalizing constant. This allows the use of classical measures such as KLD and JSD (Jensen Shannon divergence). However, despite its widespread use, its suitability for estimating the density of GANs has been questioned by Theis *et al.* [22].

Log-likelihood (or equivalently Kullback-Leibler divergence) has been the de-facto standard for training and evaluating generative models [33]. It measures the likelihood of the true data under the generated distribution on N samples from the data, *i.e.* $L = \frac{1}{N} \sum_i \log P_{\text{model}}(\mathbf{x}_i)$. Since estimating likelihood in higher dimensions is not feasible, generated samples can be used to infer something about a model’s log-likelihood. The intuition is that a model with maximum likelihood (zero KL divergence) will produce perfect samples.

The Parzen window approach to density estimation works by taking a finite set of samples generated by a model and then using those as the centroids of a Gaussian mixture. The constructed Parzen windows mixture is then used to compute a log-likelihood score on a set of test examples. Wu *et al.* [29] proposed to use annealed importance sampling (AIS) [65] to estimate log-likelihoods using a Gaussian observation model with a fixed variance. The key drawback of this approach is the assumption of the Gaussian observation model which may not work quite well in high-dimensional spaces. They found that AIS is two orders of magnitude more accurate than KDE, and is accurate enough for comparing generative models.

While likelihood is very intuitive, it suffers from several drawbacks [22]:

- (a) For a large number of samples, Parzen window estimates fall short in approximating a model’s true log-likelihood when the data dimensionality is high. Even for the fairly low dimensional space of 6×6 image patches, it requires a very large number of samples to come close to the true log-likelihood of a model. See Fig. 14.B.

¹Each sample is a vector shown in boldface (*e.g.* \mathbf{x}).

	Measure	Description
Quantitative	1. Average Log-likelihood [18, 22]	<ul style="list-style-type: none"> • Log likelihood of explaining realworld held out/test data using a density estimated from the generated data (<i>e.g.</i> using KDE or Parzen window estimation). $L = \frac{1}{N} \sum_i \log P_{model}(\mathbf{x}_i)$
	2. Coverage Metric [33]	<ul style="list-style-type: none"> • The probability mass of the true data “covered” by the model distribution $C := P_{data}(dP_{model} > t)$ with t such that $P_{model}(dP_{model} > t) = 0.95$
	3. Inception Score (IS) [3]	<ul style="list-style-type: none"> • KLD between conditional and marginal label distributions over generated data. $\exp(\mathbb{E}_{\mathbf{x}}[\text{KL}(p(y \mathbf{x}) \ p(y))])$
	4. Modified Inception Score (m-IS) [34]	<ul style="list-style-type: none"> • Encourages diversity within images sampled from a particular category. $\exp(\mathbb{E}_{\mathbf{x}_i}[\mathbb{E}_{\mathbf{x}_j}[(\text{KL}(P(y \mathbf{x}_i) \ P(y \mathbf{x}_j))])])$
	5. Mode Score (MS) [35]	<ul style="list-style-type: none"> • Similar to IS but also takes into account the prior distribution of the labels over real data. $\exp(\mathbb{E}_{\mathbf{x}}[\text{KL}(p(y \mathbf{x}) \ p(y^{train}))] - \text{KL}(p(y) \ p(y^{train})))$
	6. AM Score [36]	<ul style="list-style-type: none"> • Takes into account the KLD between distributions of training labels vs. predicted labels, as well as the entropy of predictions. $\text{KL}(p(y^{train}) \ p(y)) + \mathbb{E}_{\mathbf{x}}[H(y \mathbf{x})]$
	7. Fréchet Inception Distance (FID) [37]	<ul style="list-style-type: none"> • Wasserstein-2 distance between multi-variate Gaussians fitted to data embedded into a feature space $FID(r, g) = \ \mu_r - \mu_g\ _2^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}})$
	8. Maximum Mean Discrepancy (MMD) [38]	<ul style="list-style-type: none"> • Measures the dissimilarity between two probability distributions P_r and P_g using samples drawn independently from each distribution. $M_k(P_r, P_g) = \mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim P_r}[k(\mathbf{x}, \mathbf{x}')] - 2\mathbb{E}_{\mathbf{x} \sim P_r, \mathbf{y} \sim P_g}[k(\mathbf{x}, \mathbf{y})] + \mathbb{E}_{\mathbf{y}, \mathbf{y}' \sim P_g}[k(\mathbf{y}, \mathbf{y}')]$
	9. The Wasserstein Critic [39]	<ul style="list-style-type: none"> • The critic (<i>e.g.</i> an NN) is trained to produce high values at real samples and low values at generated samples $\hat{W}(\mathbf{x}_{test}, \mathbf{x}_g) = \frac{1}{N} \sum_{i=1}^N \hat{f}(\mathbf{x}_{test}[i]) - \frac{1}{N} \sum_{i=1}^N \hat{f}(\mathbf{x}_g[i])$
	10. Birthday Paradox Test [27]	<ul style="list-style-type: none"> • Measures the support size of a discrete (continuous) distribution by counting the duplicates (near duplicates)
	11. Classifier Two Sample Test (C2ST) [40]	<ul style="list-style-type: none"> • Answers whether two samples are drawn from the same distribution (<i>e.g.</i> by training a binary classifier)
	12. Classification Performance [1, 15]	<ul style="list-style-type: none"> • An indirect technique for evaluating the quality of unsupervised representations (<i>e.g.</i> feature extraction; FCN score). See also the GAN Quality Index (GQI) [41].
	13. Boundary Distortion [42]	<ul style="list-style-type: none"> • Measures diversity of generated samples and covariate shift using classification methods.
	14. Number of Statistically-Different Bins (NDB) [43]	<ul style="list-style-type: none"> • Given two sets of samples from the same distribution, the number of samples that fall into a given bin should be the same up to sampling noise
	15. Image Retrieval Performance [44]	<ul style="list-style-type: none"> • Measures the distributions of distances to the nearest neighbors of some query images (<i>i.e.</i> diversity)
	16. Generative Adversarial Metric (GAM) [31]	<ul style="list-style-type: none"> • Compares two GANs by having them engaged in a battle against each other by swapping discriminators or generators. $p(\mathbf{x} y=1; M_1)/p(\mathbf{x} y=1; M_2) = (p(y=1 \mathbf{x}; D_1)p(\mathbf{x}; G_2))/(p(y=1 \mathbf{x}; D_2)p(\mathbf{x}; G_1))$
	17. Tournament Win Rate and Skill Rating [45]	<ul style="list-style-type: none"> • Implements a tournament in which a player is either a discriminator that attempts to distinguish between real and fake data or a generator that attempts to fool the discriminators into accepting fake data as real.
	18. Normalized Relative Discriminative Score (NRDS) [32]	<ul style="list-style-type: none"> • Compares n GANs based on the idea that if the generated samples are closer to real ones, more epochs would be needed to distinguish them from real samples.
	19. Adversarial Accuracy and Divergence [46]	<ul style="list-style-type: none"> • Adversarial Accuracy. Computes the classification accuracies achieved by the two classifiers, one trained on real data and another on generated data, on a labeled validation set to approximate $P_g(y \mathbf{x})$ and $P_r(y \mathbf{x})$. Adversarial Divergence: Computes $\text{KL}(P_g(y \mathbf{x}), P_r(y \mathbf{x}))$
	20. Geometry Score [47]	<ul style="list-style-type: none"> • Compares geometrical properties of the underlying data manifold between real and generated data.
	21. Reconstruction Error [48]	<ul style="list-style-type: none"> • Measures the reconstruction error (<i>e.g.</i> L_2 norm) between a test image and its closest generated image by optimizing for z (<i>i.e.</i> $\min_{\mathbf{z}} \ G(\mathbf{z}) - \mathbf{x}^{(test)}\ ^2$)
	22. Image Quality Measures [49, 50, 51]	<ul style="list-style-type: none"> • Evaluates the quality of generated images using measures such as SSIM, PSNR, and sharpness difference
	23. Low-level Image Statistics [52, 53]	<ul style="list-style-type: none"> • Evaluates how similar low-level statistics of generated images are to those of natural scenes in terms of mean power spectrum, distribution of random filter responses, contrast distribution, etc.
	24. Precision, Recall and F_1 score [23]	<ul style="list-style-type: none"> • These measures are used to quantify the degree of overfitting in GANs, often over toy datasets.
Qualitative	1. Nearest Neighbors	<ul style="list-style-type: none"> • To detect overfitting, generated samples are shown next to their nearest neighbors in the training set
	2. Rapid Scene Categorization [18]	<ul style="list-style-type: none"> • In these experiments, participants are asked to distinguish generated samples from real images in a short presentation time (<i>e.g.</i> 100 ms); <i>i.e.</i> real v.s fake
	3. Preference Judgment [54, 55, 56, 57]	<ul style="list-style-type: none"> • Participants are asked to rank models in terms of the fidelity of their generated images (<i>e.g.</i> pairs, triples)
	4. Mode Drop and Collapse [58, 59]	<ul style="list-style-type: none"> • Over datasets with known modes (<i>e.g.</i> a GMM or a labeled dataset), modes are computed as by measuring the distances of generated data to mode centers
	5. Network Internals [1, 60, 61, 62, 63, 64]	<ul style="list-style-type: none"> • Regards exploring and illustrating the internal representation and dynamics of models (<i>e.g.</i> space continuity) as well as visualizing learned features

Table 1: A summary of common GAN evaluation measures.

- (b) Theis *et al.* showed that the likelihood is generally uninformative about the quality of samples and vice versa. In other words, log-likelihood and sample quality are moderately unrelated. A model can have poor log-likelihood and produce great samples, or have great log-likelihood and produce poor samples. An example in the former case is a mixture of Gaussian distributions where the means are training images (*i.e.* akin to a look-up table). Such a model will generate great samples but will still have very poor log-likelihood. An example of the latter is a mixture model combined of a good model, with a very low weight α (*e.g.* ≈ 0.01), and a bad model with a high weight $1 - \alpha$. Such a model has a large average log-likelihood but generates very poor samples (See [22] for the proof).
- (c) Parzen window estimates of the likelihood produce rankings different from other measures (See Fig. 14.C).

Due to the above issues, it becomes difficult to answer basic questions such as whether GANs are simply memorizing training examples, or whether they are missing important modes of the data distribution. For further discussions on other drawbacks of average likelihood measures consult [66].

2. **Coverage Metric.** Tolstikhin *et al.* [33] proposed to use the probability mass of the real data ‘‘covered’’ by the model distribution P_{model} as a metric. They compute $C := P_{data}(dP_{model} > t)$ with t such that $P_{model}(dP_{model} > t) = 0.95$. A kernel density estimation method was used to approximate the density of P_{model} . They claim that this metric is more interpretable than the likelihood, making it easier to assess the difference in performance of the algorithms.
3. **Inception Score (IS).** Proposed by Salimans *et al.* [3], it is perhaps the most widely adopted score for GAN evaluation (*e.g.* in [67]). It uses a pre-trained neural network (the Inception Net [68] trained on the ImageNet [69]) to capture the desirable properties of generated samples: *highly classifiable* and *diverse* with respect to class labels. It measures the average KL divergence between the conditional label distribution $p(y|\mathbf{x})$ of samples (expected to have low entropy for easily classifiable samples; better sample quality) and the marginal distribution $p(y)$ obtained from all the samples (expected to have high entropy if all classes are equally represented in the set of samples; high diversity). It favors low entropy of $p(y|\mathbf{x})$ but a large entropy of $p(y)$.

$$\exp(\mathbb{E}_{\mathbf{x}} [\text{KL}(p(y|\mathbf{x}) \| p(y))]) = \exp(H(y) - \mathbb{E}_{\mathbf{x}} [H(y|\mathbf{x})]), \quad (1)$$

where $p(y|\mathbf{x})$ is the conditional label distribution for image \mathbf{x} estimated using a pretrained Inception model [68], and $p(y)$ is the marginal distribution: $p(y) \approx 1/N \sum_{n=1}^N p(y|\mathbf{x}_n = G(\mathbf{z}_n))$. $H(\mathbf{x})$ represents entropy of variable \mathbf{x} .

The Inception score shows a reasonable correlation with the *quality* and *diversity* of generated images [3]. IS over real images can serve as the upper bound. Despite these appealing properties, IS has several limitations:

- (a) First, similar to log-likelihood, it favors a “memory GAN” that stores all training samples, thus is unable to detect overfitting (*i.e.* can be fooled by generating centers of data modes [46]). This is aggravated by the fact that it does not make use of a holdout validation set.
- (b) Second, it fails to detect whether a model has been trapped into one bad mode (*i.e.* is agnostic to mode collapse). Zhou *et al.* [36], however, shows results on the contrary.
- (c) Third, since IS uses Inception model that has been trained on ImageNet with many object classes, it may favor models that generate good objects rather realistic images.
- (d) Fourth, IS only considers P_g and ignores P_r . Manipulations such as mixing in natural images from an entirely different distribution could deceive this score. As a result, it may favor models that simply learn sharp and diversified images, instead of P_r [26]².
- (e) Fifth, it is an asymmetric measure.
- (f) Finally, it is affected by image resolution. See Fig. 2.

Zhou *et al.* [36] provide an interesting analysis of the Inception score. They experimentally measured the two components of the IS score, entropy terms in Eq. 1, during training and showed that $H(y|\mathbf{x})$ behaves as expected (*i.e.* decreasing) while $H(y)$ does not. See Fig. 3 (top row). They found that CIFAR-10 data are not evenly distributed over the classes under the Inception model trained on ImageNet. See Fig. 3(d). Using the Inception model trained over ImageNet or CIFAR-10, results in two different values for $H(y)$. Also, the value of $H(y|\mathbf{x})$ varies for each specific sample in the training data (*i.e.* some images are deemed less real than others). Further, a mode-collapsed generator usually gets a low Inception score (See Fig. 5 in [36]), which is a good sign. Theoretically, in an extreme case when all the generated samples are collapsed into a single point (thus $p(y) = p(y|\mathbf{x})$), then the minimal Inception score of 1.0 will be achieved. Despite this, it is believed that the Inception score can not reliably measure whether a model has collapsed. For example, a class-conditional model that simply memorizes one example per each ImageNet class, will achieve high IS values. Please refer to [70] for further analysis on the inception score.

4. **Modified Inception Score (m-IS).** Inception score assigns a higher value to models with a low entropy class conditional distribution over all generated data $p(y|\mathbf{x})$. However, it is desirable to have diversity within samples in a particular category. To characterize this diversity, Gurumurthy *et al.* [34] suggested to use a cross-entropy style score $-p(y|\mathbf{x}_i)\log(p(y|\mathbf{x}_j))$ where \mathbf{x}_j s are samples from the same class as \mathbf{x}_i based on the inception model’s output. Incorporating this term into the original inception-score results in:

$$\exp(\mathbb{E}_{\mathbf{x}_i}[\mathbb{E}_{\mathbf{x}_j}[(\mathbb{KL}(P(y|\mathbf{x}_i)||P(y|\mathbf{x}_j)))]], \quad (2)$$

²This also applies to the Mode Score.

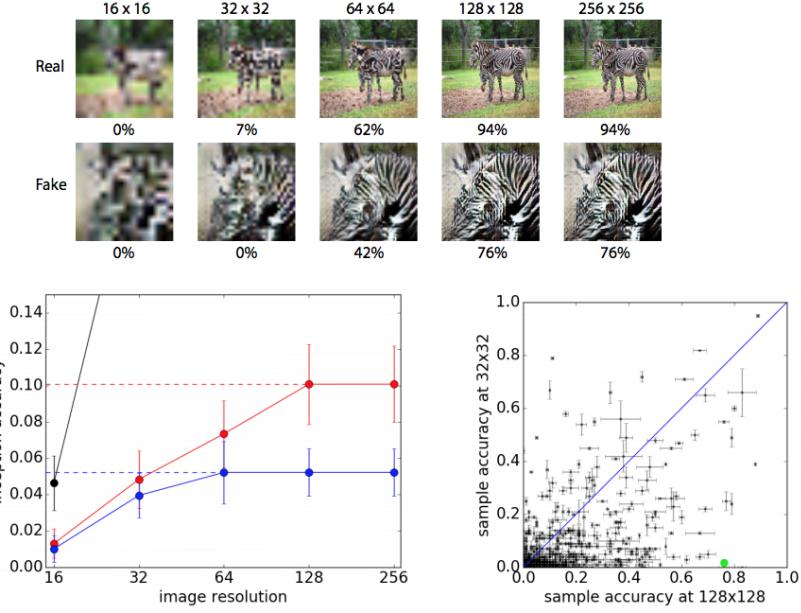


Figure 2: Sensitivity of the inception score to image resolution. Top: Training data and synthesized images from the zebra class resized to a lower spatial resolution and subsequently artificially resized to the original resolution (128×128 for the red and black lines; 64×64 for the blue line). Bottom Left: IS score across varying spatial resolutions for training data and image samples from 64×64 and 128×128 models. Error bars show standard deviation across 10 subsets of images. Dashed lines highlight the accuracy at the output spatial resolution of the model. Bottom Right: Comparison of accuracy scores at 128×128 and 32×32 spatial resolutions. Each point represents an ImageNet class. 84.4% of the classes are below the diagonal. The green dot corresponds to the zebra class. Figure from [2].

which is calculated on a per-class basis and is then averaged over all classes. Essentially, m-IS can be viewed as a proxy for measuring both intra-class sample diversity as well as sample quality.

5. **Mode Score.** Introduced in [35], this score addresses an important drawback of the Inception score which is ignoring the prior distribution of the ground truth labels (*i.e.* disregarding the dataset):

$$\exp \left(\mathbb{E}_{\mathbf{x}} [\mathbb{KL} (p(y | \mathbf{x}) \| p(y^{train}))] - \mathbb{KL} (p(y) \| p(y^{train})) \right), \quad (3)$$

where $p(y^{train})$ is the empirical distribution of labels computed from training data. Mode score adequately reflects the variety and visual quality of generated images [35]. It has been, however, proved that Inception and MODE scores are in fact equivalent. See [71] for the proof.

6. **AM Score.** Zhou *et al.* [36] argue that the entropy term on y in the Inception score is not suitable when the data is not evenly distributed over classes. To take y^{train} into account, they proposed to replace $H(y)$ with

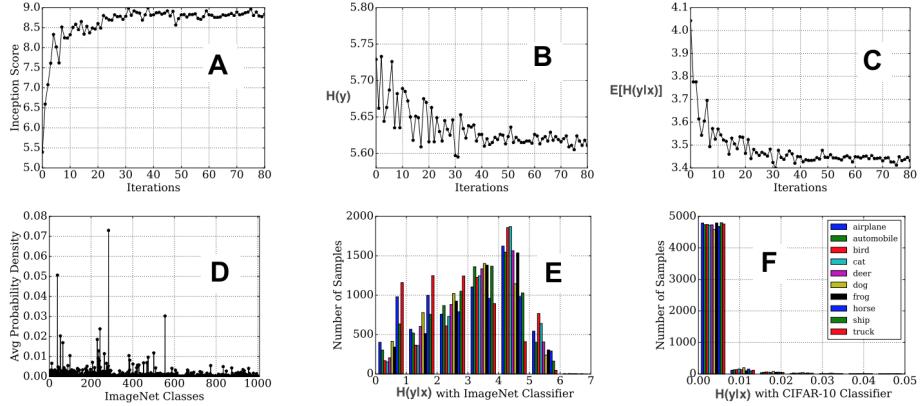


Figure 3: Top: Training curves of Inception Score and its decomposed terms. A) IS during training, B) First term in rhs of Eq. 1, $H(y)$, goes down with training which is supposed to go up, C) The second term, $H(y|\mathbf{x})$ decreases in training, as expected. Bottom: Statistics of the CIFAR-10 training images. D) $p(y)$ over ImageNet classes, E) $H(y|\mathbf{x})$ distribution with ImageNet classifier of each class, and F) $H(y|\mathbf{x})$ distribution with CIFAR-10 classifier of each class. Figure compiled from [36].

the KL divergence between y^{train} and y . The AM score is then defined as

$$\text{KL}(p(y^{train}) \parallel p(y)) + \mathbb{E}_{\mathbf{x}}[H(y|\mathbf{x})]. \quad (4)$$

The AM score consists of two terms. The first one is minimized when y^{train} is close to y . The second term is minimized when the predicted class label for sample \mathbf{x} (*i.e.* $y|\mathbf{x}$) has low entropy. Thus, the smaller the AM score, the better.

It has been shown that the Inception score with $p(y|\mathbf{x})$ being the Inception model trained with ImageNet, correlates with human evaluation on CIFAR10. CIFAR10 data, however, is not evenly distributed over the ImageNet Inception model. The entropy term on average distribution of the Inception score may thus not work well (See Fig. 3). With a pre-trained CIFAR10 classifier, the AM score can well capture the statistics of the average distribution. Thus, $p(y|\mathbf{x})$ should be a pre-trained classifier on a given dataset.

7. **Fréchet Inception Distance (FID).** Introduced by Heusel *et al.* [37], FID embeds a set of generated samples into a feature space given by a specific layer of Inception Net (or any CNN). Viewing the embedding layer as a continuous multivariate Gaussian, the mean and covariance are estimated for both the generated data and the real data. The Fréchet distance between these two Gaussians (*a.k.a* Wasserstein-2 distance) is then used to quantify the quality of generated samples, *i.e.*,

$$FID(r, g) = \|\mu_r - \mu_g\|_2^2 + Tr \left(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}} \right), \quad (5)$$

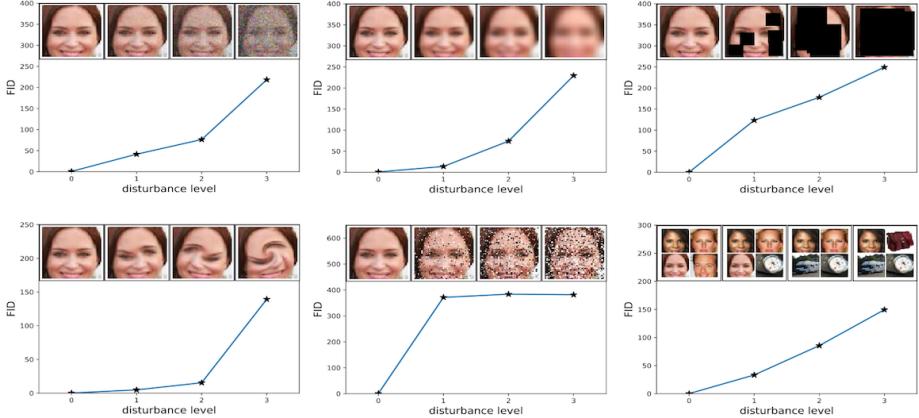


Figure 4: FID measure is sensitive to image distortions. From upper left to lower right: Gaussian noise, Gaussian blur, implanted black rectangles, swirled images, salt and pepper noise, and CelebA dataset contaminated by ImageNet images. Figure from [37].

where (μ_r, Σ_r) and (μ_g, Σ_g) are the mean and covariance of the real data and model distributions, respectively. Lower FID means smaller distances between synthetic and real data distributions.

FID performs well in terms of discriminability, robustness and computational efficiency. It appears to be a good measure, even though it only takes into consideration the first two order moments of the distributions. However, it assumes that features are of Gaussian distribution which is often not guaranteed. It has been shown that FID is consistent with human judgments and is more robust to noise than IS [37] (e.g. negative correlation between the FID and visual quality of generated samples). Unlike IS however, it is able to detect intra-class mode dropping³, i.e. a model that generates only one image per class can score a high IS but will have a bad FID. Also, unlike IS, the FID worsens as various types of artifacts are added to images (See Fig. 4). IS and AM scores measure the diversity and quality of generated samples, while FID measures the distance between the generated and real distributions. An empirical analysis of FID can be found in [23]. See also [73] for a class-aware version of FID.

8. **Maximum Mean Discrepancy (MMD).** This measure computes the dissimilarity between two probability distributions⁴ P_r and P_g using samples drawn independently from each [74]. A lower MMD hence means that P_g is closer to P_r . MMD can be regarded as two-sample testing since, as in classifier two samples test, it tests whether one model or another is closer to the true data distribution [75, 76, 77]. Such hypothesis tests

³On the contrary, Sajjadi et al. [72] show that FID is sensitive to both the addition of spurious modes as well as to mode dropping.

⁴Distinguishing two distributions by finite samples is known as *Two-Sample Test* in statistics.

allow choosing one evaluation measure over another.

The kernel MMD [38] measures (square) MMD between P_r and P_g for some fixed characteristic kernel function k (*e.g.* Gaussian kernel $k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2)$) as follows⁵:

$$M_k(P_r, P_g) = \mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim P_r} [k(\mathbf{x}, \mathbf{x}')] - 2\mathbb{E}_{\mathbf{x} \sim P_r, \mathbf{y} \sim P_g} [k(\mathbf{x}, \mathbf{y})] + \mathbb{E}_{\mathbf{y}, \mathbf{y}' \sim P_g} [k(\mathbf{y}, \mathbf{y}')]. \quad (6)$$

In practice, finite samples from distributions are used to estimate MMD distance. Given $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \sim P_r$ and $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_n\} \sim P_g$, one estimator of $M_k(P_r, P_g)$ is:

$$\hat{M}_k(X, Y) = \frac{1}{\binom{n}{2}} \sum_{i \neq i'} k(\mathbf{x}_i, \mathbf{x}'_{i'}) - \frac{2}{\binom{n}{2}} \sum_{i \neq j} k(\mathbf{x}_i, \mathbf{y}_j) + \frac{1}{\binom{n}{2}} \sum_{j \neq j'} k(\mathbf{y}_j, \mathbf{y}'_{j'}). \quad (7)$$

Because of the sampling variance, $\hat{M}(X, Y)$ may not be zero even when $P_r = P_g$. Li *et al.* [78] put forth a remedy to address this. Kernel MMD works surprisingly well when it operates in the feature space of a pre-trained CNN. It is able to distinguish generated images from real images, and both its sample complexity and computational complexity are low [26].

Kernel MMD has also been used for training GANs. For example, the Generative Moment Matching Network (GMMN) [79, 80, 78] replaces the discriminator in GAN with a two-sample test based on kernel MMD. See also [81] for more analyses on MMD and its use in GAN training.

9. **The Wasserstein Critic.** The Wasserstein critic [39] provides an approximation of the Wasserstein distance between the real data distribution P_r and the generator distribution P_g :

$$W(P_r, P_g) \propto \max_f \mathbb{E}_{\mathbf{x} \sim P_r} [f(\mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim P_g} [f(\mathbf{x})], \quad (8)$$

where $f : \mathbb{R}^D \rightarrow \mathbb{R}$ is a Lipschitz continuous function. In practice, the critic f is a neural network with clipped weights to have bounded derivatives. It is trained to produce high values at real samples and low values at generated samples (*i.e.* is an approximation):

$$\hat{W}(\mathbf{x}_{test}, \mathbf{x}_g) = \frac{1}{N} \sum_{i=1}^N \hat{f}(\mathbf{x}_{test}[i]) - \frac{1}{N} \sum_{i=1}^N \hat{f}(\mathbf{x}_g[i]) \quad (9)$$

where \mathbf{x}_{test} is a batch of samples from a test set, \mathbf{x}_g is a batch of generated samples, and \hat{f} is the independent critic. For discrete distributions with densities P_r and P_g , the Wasserstein distance is often referred to as the Earth Mover's Distance (EMD) which intuitively is the minimum mass displacement to transform one distribution into the other. A variant of this score known as sliced Wasserstein distance (SWD) approximates the

⁵Please beware that here \mathbf{y} represents the generated samples, and not the class labels.

Wasserstein-1 distance between real and generated images, and is computed as the statistical similarity between local image patches extracted from Laplacian pyramid representations of these images [53]. We will discuss SWD in more detail later under scores that utilize low-level image statistics. This measure addresses both overfitting and mode collapse. If the generator memorizes the training set, the critic trained on test data can distinguish between samples and data. If mode collapse occurs, the critic will have an easy job in distinguishing between data and samples. Further, it does not saturate when the two distributions do not overlap. The magnitude of the distance indicates how easy it is for the critic to distinguish between samples and data.

The Wasserstein distance works well when the base distance is computed in a suitable feature space. A key limitation of this distance is its high sample and time complexity. These make Wasserstein distance less appealing as a practical evaluation measure, compared to other ones (See [27]).

10. **Birthday Paradox Test.** This test approximates the support size⁶ of a discrete distribution. Arora and Zhang [27] proposed to use the birthday paradox⁷ test to evaluate GANs as follows:

- (a) Pick a sample of size S from the generated distribution
- (b) Use an automated measure of image similarity to flag the k (*e.g.* $k = 20$) most similar pairs in the sample
- (c) Visually inspect the flagged pairs and check for duplicates
- (d) Repeat.

The suggested plan is to manually check for duplicates in a sample of size S . If a duplicate exists, then the estimated support size is S^2 . It is not possible to find exact duplicates as the distribution of generated images is continuous. Instead, a distance measure can be used to find near-duplicates (*e.g.* using the L_2 norm). In practice, they first created a candidate pool of potential near-duplicates by choosing the 20 closest pairs according to some heuristic measure, and then visually identified the near duplicated. Following this procedure and using Euclidean distance in pixel space, Arora and Zhang [27] found that with probability $\geq 50\%$, a batch of about 400 samples generated from the CelebA dataset [82] contains at least one pair of duplicates for both DCGAN and MIX+DCGAN (thus leading to support size of 400^2). The birthday theorem assumes uniform sampling. Arora and Zhang [27], however, claim that the birthday paradox holds even if data are distributed in a highly nonuniform way. This test can be used to detect mode collapse in GANs.

11. **Classifier Two-sample Tests (C2ST).** The goal of two-sample tests is to assess whether two samples are drawn from the same distribution [40].

⁶The support of a real-valued function f is the subset of the domain containing those elements which are not mapped to zero.

⁷The “Birthday theorem” states that with probability at least 50%, a uniform sample (with replacement) of size S from a set of N elements will have a duplicate given $S > \sqrt{N}$.

In other words, decide whether two probability distributions, denoted by P and Q , are equal. The generator is evaluated on a held out test set. This set is split into a test-train and test-test subsets. The test-train set is used to train a fresh discriminator, which tries to distinguish generated images from the real images. Afterwards, the final score is computed as the performance of this new discriminator on the test-test set and the freshly generated images. More formally, assume we have access to two samples $S_P = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \sim P^n(X)$ and $S_Q = \{\mathbf{y}_1, \dots, \mathbf{y}_n\} \sim Q^n(Y)$ where $\mathbf{x}_i, \mathbf{y}_i \in \mathcal{X}$, for all $i = 1, \dots, n$. To test whether the null hypothesis $H_0 : P = Q$ is true, these five steps need to be completed:

- (a) Construct the following dataset

$$\mathcal{D} = \{(\mathbf{x}_i, 0)\}_{i=1}^n \cup \{(\mathbf{y}_i, 1)\}_{i=1}^n =: \{(\mathbf{z}_i, l_i)\}_{i=1}^{2n}.$$

- (b) Randomly shuffle \mathcal{D} , and split it into two disjoint *training* and *testing* subsets $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$, where $\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{test}}$ and $n_{\text{test}} := |\mathcal{D}_{\text{test}}|$.
- (c) Train a binary classifier $f : \mathcal{X} \rightarrow [0, 1]$ on $\mathcal{D}_{\text{train}}$. In the following, assume that $f(\mathbf{z}_i)$ is an estimate of the conditional probability distribution $p(l_i = 1 | \mathbf{z}_i)$.
- (d) Calculate the classification accuracy on $\mathcal{D}_{\text{test}}$:

$$\hat{t} = \frac{1}{n_{\text{test}}} \sum_{(\mathbf{z}_i, l_i) \in \mathcal{D}_{\text{test}}} \mathbb{I} \left[\mathbb{I} \left(f(\mathbf{z}_i) > \frac{1}{2} \right) = l_i \right] \quad (10)$$

as the *C2ST statistic*, where \mathbb{I} is the indicator function. The intuition here is that if $P = Q$, the test accuracy in Eq. 10 should remain near chance-level. In contrast, if binary classifier performs better than chance then it implies that $P \neq Q$.

- (e) To accept or reject the null hypothesis, compute a p -value using the null distribution of the C2ST.

In principle, any binary classifier can be adopted for computing C2ST. Huang *et al.* [26] introduce a variation of this measure known as the *1-Nearest Neighbor* classifier. The advantage of using 1-NN over other classifiers is that it requires no special training and little hyperparameter tuning. Given two sets of real S_r and generated S_g samples with the same size (*i.e.* $|S_r| = |S_g|$), one can compute the leave-one-out (LOO) accuracy of a 1-NN classifier trained on S_r and S_g with positive labels for S_r and negative labels for S_g . The LOO accuracy can vary from 0% to 100%. If the GAN memorizes samples in S_r and re-generate them exactly, *i.e.* $S_g = S_r$, then the accuracy would be 0%. This is because every sample from S_r would have its nearest neighbor from S_g with zero distance (and vice versa). If it generates samples that are widely different than real images (and thus completely separable), then the performance would be 100%. Notice that chance level here is 50% which happens when a label is randomly assigned to an image. Lopez-Paz and Oquab [83] offer a revisit of classifier two-sample tests in [83].

Classifier two-sample tests can be considered a different form of two-sample test to MMD. MMD has the advantage of a U-statistic estimator with Gaussian asymptotic distribution, while the classifier 2-sample test has a different form (cf. [75]). MMD can be better when the U-statistic convergence outweighs the potentially more powerful classifier (*e.g.* from a deep network), while a classifier based test could be better if the classifier is better than the choice of kernel.

12. **Classification Performance.** One common indirect technique for evaluating the quality of unsupervised representation learning algorithms is to apply them as feature extractors on labeled datasets and evaluate the performance of linear models fitted on top of the learned features. For example, to evaluate the quality of the representations learned by DC-GANs, Radford *et al.* [1] trained their model on ImageNet dataset and then used the discriminator’s convolutional features from all layers to train a regularized linear L2-SVM to classify CIFAR-10 images. They achieved 82.8% accuracy on par with or better than several baselines trained directly on CIFAR-10 data.

A similar strategy has also been followed in evaluating conditional GANs (*e.g.* the ones proposed for style transfer). For example, an off-the-shelf classifier is utilized by Zhang *et al.* [84] to assess the realism of synthesized images. They fed their fake colorized images to a VGG network that was trained on real color photos. If the classifier performs well, this indicates that the colorizations are accurate enough to be informative about object class. They call this “semantic interpretability”. Similarly, Isola *et al.* [15] proposed the “FCN score” to measure the quality of the generated images conditioned on an input segmentation map. They fed the generated images to the fully-convolutional semantic segmentation network (FCN) [85] and then measured the error between the output segmentation map and the ground truth segmentation mask.

Ye et al. [41] proposed an objective measure known as the **GAN Quality Index (GQI)** to evaluate GANs. First, a generator G is trained on a labeled real dataset with N classes. Next, a classifier C_{real} is trained on the real dataset. The generated images are then fed to this classifier to obtain labels. A second classifier, called the GAN-induced classifier C_{GAN} , is trained on the generated data. Finally, the GQI is defined as the ratio of the accuracies of the two classifiers:

$$GQI = \frac{ACC(C_{GAN})}{ACC(C_{real})} \times 100 \quad (11)$$

GQI is an integer in the range of 0 to 100. Higher GQI means that the GAN distribution better matches the real data distribution.

Data Augmentation Utility: Some works measure the utility of GANs for generating additional training samples. This can be interpreted as a measure of the diversity of the generated images. Similar to Ye et al. [41], Lesort et al. [86] proposed to use a mixture of real and generated data to train a classifier and then test it on a labeled test dataset. The result is

then compared with the score of the same classifier trained on the real training data mixed with noise. Along the same line, recently, Shmelkov et al. [25] proposed to compare class-conditional GANs with GAN-train and GAN-test scores using a neural net classifier. GAN-train is a network trained on GAN generated images and is evaluated on real-world images. GAN-test, on the other hand, is the accuracy of a network trained on real images and evaluated on the generated images. They analyzed the diversity of the generated images by evaluating GAN-train accuracy with varying amounts of generated data. The intuition is that a model with low diversity generates redundant samples, and thus increasing the quantity of data generated in this case does not result in better GAN-train accuracy. In contrast, generating more samples from a model with high diversity produces a better GAN-train score.

Above mentioned measures are indirect and rely heavily on the choice of the classifier. Nonetheless, they are useful for evaluating generative models based on the notion that a better generative model should result in better representations for surrogate tasks (*e.g.* supervised classification). This, however, does not necessary imply that generated images have high diversity.

13. **Boundary Distortion.** Santurkar et al. [42] aimed to measure diversity of generated samples using classification methods. This phenomenon can be viewed as a form of covariate shift in GANs wherein the generator concentrates a large probability mass on a few modes of the true distribution. It is illustrated using two toy examples in Fig. 5. The first example regards learning a unimodal spherical Gaussian distribution using a vanilla GAN [18]. As can be seen in Fig. 5.A, the spectrum (eigenvalues of the covariance matrix) of GAN data shows a decaying behavior (unlike true data). The second example considers binary classification using logistic regression where the true distribution for each class is a unimodal spherical Gaussian. The synthetic distribution for one of the classes undergoes boundary distortion which causes a skew between the classifiers trained on true and synthetic data (Fig. 5.B). Naturally, such errors would lead to poor generalization performance on true data as well. Taken together, these examples show that a) boundary distortion is a form of covariate shift that GANs can realistically introduce, and b) this form of diversity loss can be detected and quantified even using classification.

Specifically, Santurkar et al. proposed the following method to measure boundary distortion introduced by a GAN:

- (a) Train two separate instances of the given unconditional GAN, one for each class in true dataset D (assume two classes).
- (b) Generate a balanced dataset by drawing $N/2$ from each of these GANs.
- (c) Train a binary classifier based on the labeled GAN dataset obtained in Step 2 above.
- (d) Train an identical, in terms of architecture and hyperparameters, classifier on the true data D for comparison.

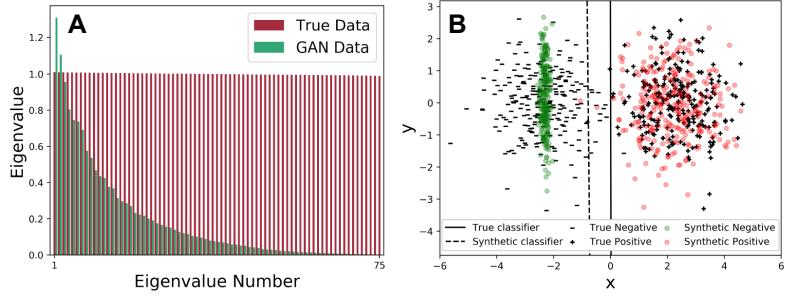


Figure 5: A) Spectrum of the learned distribution of a vanilla GAN (a 200D latent space), compared to that of the true distribution (a 75D spherical unimodal Gaussian). B) An example of covariate shift between synthetic and true distributions leading to a distortion in the learned decision boundary of a (linear) logistic regression classifier. Here the synthetic distribution for one class suffers from boundary distortion. Figure compiled from [32].

Afterwards, the performance of both classifiers is measured on a hold-out set of true data. Performance of the classifier trained on synthetic data on this set acts as a proxy measure for diversity loss through covariate shift. Notice that this measure is akin to the classification performance discussed above.

14. **Number of Statistically-Different Bins (NDB).** To measure diversity of generated samples and mode collapse, Richardson and Weiss [43] propose an evaluation method based on the following observation: Given two sets of samples from the same distribution, the number of samples that fall into a given bin should be the same up to sampling noise. More formally, let $I_B(\mathbf{x})$ be the indicator function for bin B . $I_B(\mathbf{x}) = 1$ if the sample \mathbf{x} falls into the bin B and zero otherwise. Let $\{\mathbf{x}_i^p\}$ be N_p samples from distribution p (e.g. training samples) and $\{\mathbf{x}_j^q\}$ be N_q samples from distribution q (e.g. testing samples), then if $p = q$, it is expected that $\frac{1}{N_p} \sum_i I_B(\mathbf{x}_i^p) \approx \frac{1}{N_q} \sum_j I_B(\mathbf{x}_j^q)$. The *pooled sample proportion* P (the proportion that falls into B in the joined sets) and its standard error: $SE = \sqrt{P(1-P)[1/N_p + 1/N_q]}$ are calculated. The test statistic is the z -score: $z = \frac{P_p - P_q}{SE}$, where P_p and P_q are the proportions from each sample that fall into bin B . If z is smaller than a threshold (i.e. significance level) then the number is *statistically different*. This test is performed on all bins and then the *number of statistically-different bins* (NDB) is reported.
- To perform binning, one option is to use a uniform grid. The drawback here is that in high dimensions, a randomly chosen bin in a uniform grid is very likely to be empty. Richardson and Weiss proposed to use *Voronoi cells* to guarantee that each bin will contain some samples. Fig. 6 demonstrates this procedure using a toy example in \mathbb{R}^2 . To define the Voronoi cells, N_p training samples are clustered into K ($K \ll N_p, N_q$) clusters using K-means. Each training sample \mathbf{x}_i^p is assigned to one of the K cells (bins). Each generated sample \mathbf{x}_j^q is then assigned to the nearest (L_2) of the K

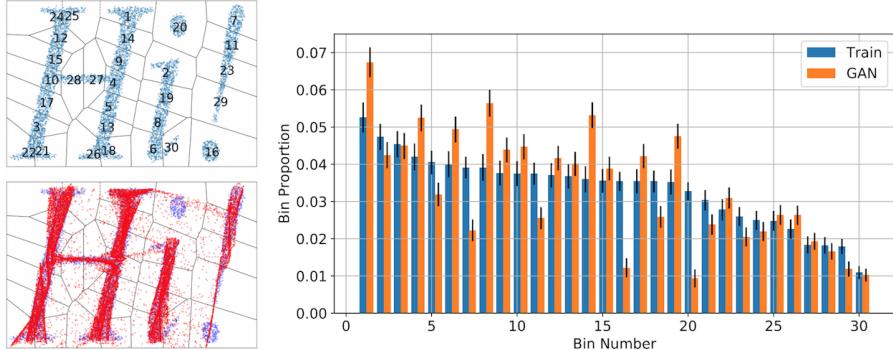


Figure 6: Illustration of the NDB evaluation method on a toy example in \mathbb{R}^2 . Top-left: The training data (blue) and binning result - Voronoi cells (numbered by bin size). Bottom-left: Samples (red) drawn from a GAN trained on the data. Right: Comparison of bin proportions between the training data and the GAN samples. Black lines = standard error (SE) values. Figure from [43].

centroids.

Unlike IS and FID, NDB measure is applied directly on the image pixels rather than pre-learned deep representations. This makes NDB domain agnostic and sensitive to different image artifacts (as opposed to using pre-trained deep models). One advantage of NDB over MS-SSIM and Birthday Paradox Test is that NDB offers a measure between the data and generated distributions and not just measuring the general diversity of the generated samples. One concern regarding NDB is that using L_2 distance in pixel space as a measure of similarity may not be meaningful.

15. **Image Retrieval Performance.** Wang *et al.* [44] proposed an image retrieval measure to evaluate GANs. The main idea is to investigate images in the dataset that are badly modeled by a network. Images from a held-out test set as well as generated images are represented using a discriminatively trained CNN [87]. The nearest neighbors of generated images in the test dataset are then retrieved. To evaluate the quality of the retrieval results, they proposed two measures:

- (a) Measure 1: Consider $d_{i,j}^k$ to be the distance of the j^{th} nearest image generated by method k to test image i , and $\mathbf{d}_j^k = \{d_{1,j}^k, \dots, d_{n,j}^k\}$ the set of j^{th} -nearest distances to all n test images (j is often set to 1). The Wilcoxon signed-rank test is then used to test the hypothesis that the median of the difference between two nearest distance distributions by two generators is zero, in which case they are equally good (*i.e.* the median of the distribution $\mathbf{d}_1^k - \mathbf{d}_1^m$). If they are not equal, the test can be used to assess which method is statistically better.
- (b) Measure 2: Consider \mathbf{d}_j^t to be the distribution of the j^{th} nearest distance of the train images to the test dataset. Since train and test sets are drawn from the same dataset, the distribution \mathbf{d}_j^t can

be considered the optimal distribution that a generator could attain (assuming it generates an equal number of images present in the train set). To model the difference with this ideal distribution, the relative increase in mean nearest neighbor distance is computed as:

$$\hat{d}_j^k = \frac{\bar{d}_j^k - \bar{d}_j^t}{\bar{d}_j^t}, \quad \bar{d}_j^k = \frac{1}{N} \sum_{i=1}^N d_{i,j}^k, \quad \bar{d}_j^t = \frac{1}{N} \sum_{i=1}^N d_{i,j}^t, \quad (12)$$

where N is the size of the test dataset. As an example, $\hat{d}_1 = 0.1$ for a model means that the average distance to the nearest neighbor of a query image is 10% higher than for data drawn from the real distribution.

16. **Generative Adversarial Metric (GAM).** Im *et al.* [31] proposed to compare two GANs by having them engaged in a battle against each other by swapping discriminators or generators across the two models (See Fig. 7). GAM measures the relative performance of two GANs by measuring the likelihood ratio of the two models. Consider two GANs with their respective trained partners, $M_1 = (D_1, G_1)$ and $M_2 = (D_2, G_2)$, where G_1 and G_2 are the generators, and D_1 and D_2 are the discriminators. The hypothesis $H1$ is that M_1 is better than M_2 if G_1 fools D_2 more than G_2 fools D_1 , and vice versa for the hypothesis $H0$. The likelihood-ratio is defined as:

$$\frac{p(\mathbf{x}|y=1; M'_1)}{p(\mathbf{x}|y=1; M'_2)} = \frac{p(y=1|\mathbf{x}; D_1)p(\mathbf{x}; G_2)}{p(y=1|\mathbf{x}; D_2)p(\mathbf{x}; G_1)}, \quad (13)$$

where M'_1 and M'_2 are the swapped pairs (D_1, G_2) and (D_2, G_1) , $p(\mathbf{x}|y=1; M)$ is the likelihood of \mathbf{x} generated from the data distribution $p(\mathbf{x})$ by model M , and $p(y=1|\mathbf{x}; D)$ indicates that discriminator D thinks \mathbf{x} is a real sample.

Then, one can measure which generator fools the opponent's discriminator more, $\frac{D_1(\mathbf{x}_2)}{D_2(\mathbf{x}_1)}$ where $\mathbf{x}_1 \sim G_1$ and $\mathbf{x}_2 \sim G_2$. To do so, Im *et al.* proposed a sample ratio test to declare a winner or a tie.

A variation of GAM known as generative multi-adversarial metric (GMAM), that is amenable to training with multiple discriminators, has been proposed in [88].

GAM suffers from two main caveats: a) it has a constraint where the two discriminators must have an approximately similar performance on a calibration dataset, which can be difficult to satisfy in practice, and b) it is expensive to compute because it has to be computed for all pairs of models (*i.e.* pairwise comparisons between independently trained GAN).

17. **Tournament Win Rate and Skill Rating.** Inspired by GAM and GMAM scores (mentioned above) as well as skill rating systems in games such as chess or tennis, Olsson *et al.* [45] utilized tournaments between generators and discriminators for GAN evaluation. They introduced two methods for summarizing tournament outcomes: tournament win rate

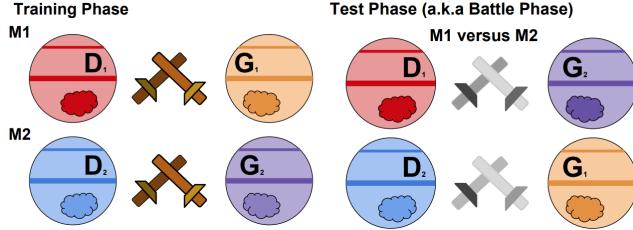


Figure 7: Illustration of the Generative Adversarial Metric (GAM). During the training phase, G_1 and G_2 compete with D_1 and D_2 , respectively. At test time, model M1 plays against M2 by having G_1 try to fool D_2 , and vice-versa. M1 is better than M2 if G_1 fools D_2 more than G_2 fools D_1 (and vice versa). Figure from [31].

and skill rating. Evaluations are useful in different contexts, including a) monitoring the progress of a single model as it learns during the training process, and b) comparing the capabilities of two different fully trained models. The former regards a single model playing against past and future versions of itself producing a useful measure of training progress (*a.k.a* within trajectory tournament). The latter regards multiple separate models (using different seeds, hyperparameters, and architectures) and provides a useful relative comparison between different trained GANs (*a.k.a* multiple trajectory tournament). Each player in a tournament is either a discriminator that attempts to distinguish between real and fake data or a generator that attempts to fool the discriminators into accepting fake data as real.

Tournament Win Rate: To determine the outcome of a match between discriminator D and generator G , the discriminator D judges two batches: one batch of samples from generator G , and one batch of real data. Every sample \mathbf{x} that is not judged correctly by the discriminator (*e.g.* $D(\mathbf{x}) \geq 0.5$ for the generated data or $D(\mathbf{x}) \leq 0.5$ for the real data) counts as a win for the generator and is used to compute its win rate. A match win rate of 0.5 for G means that D 's performance against G is no better than chance. The tournament win rate for generator G is computed as its average win rate over all discriminators in D . Tournament win rates are interpretable only within the context of the tournament they were produced from, and cannot be directly compared with those from other tournaments.

Olsson et al. ran a tournament between 20 saved checkpoints of discriminators and generators from the same training run of a DCGAN trained on SVHN [89] using an evaluation batch size of 64. Fig. 8.A shows the raw tournament outcomes from the within-trajectory tournament, alongside the same tournament outcomes summarized using tournament win rate and skill rating, as well as SVHN classifier score and SVHN Fréchet distance computed from 10,000 samples, for comparison⁸. It shows that tournament

⁸To compute these scores, a pre-trained SVHN classifier is used rather than an ImageNet

win rate and skill rating both provide a comparable measure of training progress to SVHN classifier score.

Skill Rating: Here the idea is to use a skill rating system to summarize tournament outcomes in a way that takes into account the amount of new information each match provides. Olsson et al. used the Glicko2 system [90]. In a nutshell, a player’s skill rating is represented as a Gaussian distribution, with a mean and standard deviation, representing the current state of the evidence about their “true” skill rating. See [45] for details of the algorithm. Olsson et al. constructed a tournament from saved snapshots from six SVHN GANs that differ slightly from one another, including different loss functions and architectures. They included 20 saved checkpoints of discriminators and generators from each GAN experiment, a single snapshot of 6-auto, and a generator player that produces batches of real data as a benchmark. Fig. 8.B shows the results compared to Inception and Fréchet distances.

One advantage of these scores is that they are not limited to fixed feature sets and players can learn to attend to any features that are useful to win. Another advantage is that human judges are eligible to play as discriminators, and could participate to receive a skill rating. This allows a principled method to incorporate human perceptual judgments in model evaluation. The downside is providing relative rather than absolute score of a model’s ability thus making reproducing results challenging and expensive.

18. **Normalized Relative Discriminative Score (NRDS).** The main idea behind this measure proposed by Zhang *et al.* [32] is that more epochs would be needed to distinguish good generated samples from real samples (compared to separating poor ones from real samples). They used a binary classifier (discriminator) to separate the real samples from fake ones generated by all the models in comparison. In each training epoch, the discriminator’s output for each sample is recorded. The average discriminator output of real samples will increase with epoch (approaching 1), while that of generated samples from each model will decrease (approaching 0). However, the decrement rate of each model varies based on how close the generated samples are to the real ones. The samples closer to real ones show slower decrement rate whereas poor samples will show a faster decrement rate. Therefore, comparing the “decrement rate” of each model can be an indication of how well it performs relative to other models.

There are three steps to compute the NRDS:

- (a) Obtain the curve \mathcal{C}_i ($i = 1, 2, \dots, n$) of discriminator average output versus epoch (or mini-batch) for each model (assuming n models in comparison) during training,
- (b) Compute the area under each curve $A(\mathcal{C}_i)$ (as the decrement rate), and

classifier.

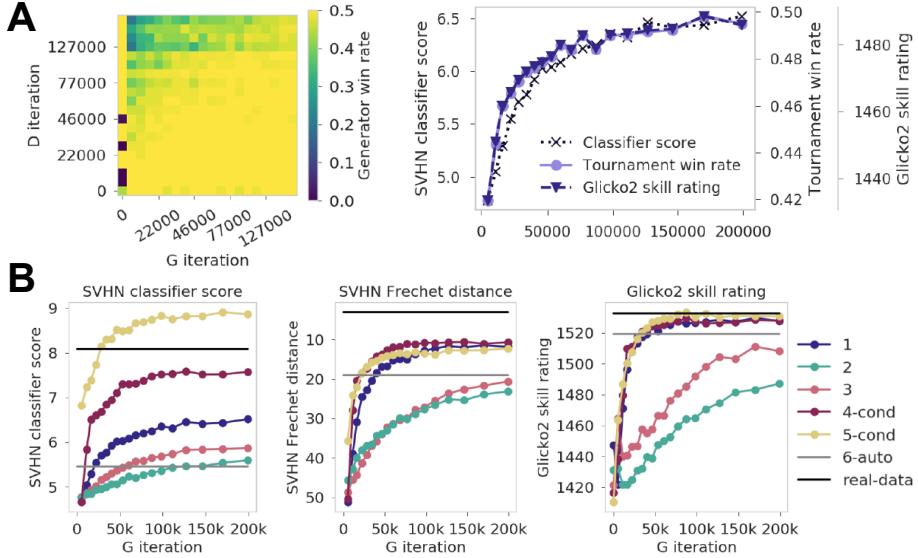


Figure 8: A) A within-trajectory tournament. Left panel shows raw tournament outcomes. Each pixel represents the average win rate between one generator and one discriminator from different iterations. Brighter pixel values represent stronger generator performance. Right panel compares tournament summary measures to SVHN classifier score. Tournament win rate in this figure is the column-wise average of the pixel values in the heatmap. B) Multiple-trajectory tournament outcomes among six models and real data. The tournament contains SVHN generator and discriminator snapshots from models with different seeds, hyperparameters, and architectures. Models are evaluated using SVHN classifier score (left), SVHN Fréchet distance (center), and skill rating method (right). Each point represents the score of one iteration of one model. The overall trajectories show the improvement of each model with increasing training. Note that the inverted y-axis on the Fréchet distance plot, such that lower distance (better quality) is plotted higher on the plot. The learning curves produced by skill rating broadly agree with those produced by Fréchet distance, and disagree with classifier score only in the case of the conditional models 4-cond and 5-con. Figure compiled from [45]. Please see text for more details on these experiments.

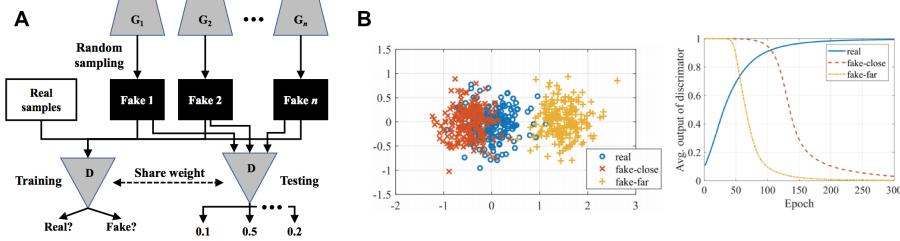


Figure 9: A) Illustration of NRDS. G_n indicates the n th generative model. Its corresponding fake samples are Fake n , which are sampled randomly. The fake samples from n models, as well as the real samples, are used to train the binary classifier D during training (bottom left). Testing only uses fake samples and performs alternatively with the training process. The bottom right shows an example of average output of D for fake samples of each model. B) A toy example of computing NRDS. Left: the real and fake samples randomly sampled from 2D normal distributions with different means but with the same (identity) covariance. The real samples (blue circles) have with zero mean. The red “x” and yellow “+” denote fake samples with the mean of [0.5, 0] and [1.5, 0], respectively. The notation fake-close (fake-far) indicates that the mean of correspondingly fake samples is close to (far from) that of the real samples. Right: the curves of epoch vs. averaged output of discriminator on corresponding sets (colors) of samples. In this example, the area under the curves of fake-close (C_1) and fake-far (C_2) are $A(C_1) = 145.4955$ and $A(C_2) = 71.1057$, respectively. From Eq. 14, $NRDS_1 = \frac{A(C_1)}{\sum_{i=1}^n A(C_i)} = 0.6717$ and $NRDS_2 = \frac{A(C_2)}{\sum_{i=1}^n A(C_i)} = 0.3283$. Therefore, the model generating fake-close is relatively better. Figure compiled from [32].

(c) Compute NRDS of the i th model by

$$NRDS_i = \frac{A(\mathcal{C}_i)}{\sum_{j=1}^n A(\mathcal{C}_j)}. \quad (14)$$

The higher the NRDS, the better. Fig. 9 illustrates the computation of NRDS over a toy example.

19. **Adversarial Accuracy and Adversarial Divergence.** Yang *et al.* [46]

proposed two measures based on the intuition that a sufficient, but unnecessary, condition for closeness of generated data distribution $P_g(\mathbf{x})$ and the real data distribution $P_r(\mathbf{x})$ is closeness of $P_g(\mathbf{x}|y)$ and $P_r(\mathbf{x}|y)$, *i.e.* distributions of generated data and real data conditioned on all possible variables of interest y , *e.g.* category labels. One way to obtain the variable of interest y is by asking human participants to annotate the images (sampled from $P_g(\mathbf{x})$ and $P_r(\mathbf{x})$).

Since it is not feasible to directly compare $P_g(\mathbf{x}|y)$ and $P_r(\mathbf{x}|y)$, they proposed to compare $P_g(y|\mathbf{x})$ and $P_r(y|\mathbf{x})$ instead (following the Bayes rule) which is a much easier task. Two classifiers are then trained from human annotations to approximate $P_g(y|\mathbf{x})$ and $P_r(y|\mathbf{x})$ for different categories. These classifiers are used to compute the following evaluation measures:

- (a) *Adversarial Accuracy:* Computes the classification accuracies achieved by the two classifiers on a validation set (*i.e.* another set of real images). If $P_g(\mathbf{x})$ is close to $P_r(\mathbf{x})$, then similar accuracies are expected.

- (b) *Adversarial Divergence*: Computes the KL divergence between $P_g(y|\mathbf{x})$ and $P_r(y|\mathbf{x})$. The lower the adversarial divergence, the closer the two distributions. The lower bound for this measure is exactly zero, which means $P_g(y|\mathbf{x}) = P_r(y|\mathbf{x})$ for all samples in the validation set.

One drawback of these measures is that a lot of human effort is needed to label the real and generated samples. To mitigate this, Yang *et al.* [46] first trained one generator per category using a labeled training set and then generated samples from all categories. Notice that these measures overlap with classification performance discussed above.

20. **Geometry Score.** Khrulkov and Oseledets [47] proposed to compare geometrical properties of the underlying data manifold between real and generated data. This score, however, involves a lot of technical details making it hard to understand and compute. Here, we provide an intuitive description.

The core idea is to build a simplicial complex from data using proximity information (*e.g.* pairwise distances between samples). To investigate the structure of the manifold, a threshold ϵ is varied and generated simplices are added into the approximation. An example is shown in Fig. 10.A. For each value of ϵ , topological properties of the corresponding simplicial complex, namely homologies, are computed. A homology encodes the number of holes of various dimensions in a space. Eventually, a barcode (signature) is constructed reflecting how long generated holes (homologies) persist in simplicial complexes (Fig. 10.B). In general, to find the rank of a k -homology (*i.e.* the number of k -dimensional holes) at some fixed value ϵ , one has to count intersections of the vertical line $\epsilon = 0$ with the intervals at the desired block H_k .

Since computing the barcode using all data is intractable, in practice often subsets of data (*e.g.* by randomly selecting points) are used. For each subset, Relative Living Times (RLT) of each number of holes is computed which is defined as the ratio of the total time when this number was present and of the value ϵ_{max} when points connect into a single blob. The RLT over random subsets are then averaged to give the Mean Relative Living Times (MRLT). By construction, they add up to 1. To quantitatively evaluate the topological difference between two datasets, the L_2 distance between these distributions is computed.

Fig. 10.C shows an example over synthetic data. Intuitively, the value at location i in the bar chart (on x axis), indicates that for that amount of time, the 1D hole existed by varying the threshold. For example, in the left most histogram, nearly never none, 2 or 3 1D holes were observed and most of the time only one hole appeared. Similarly, for the 4th pattern from the left, most of the time one 1D hole is observed. Comparing the MRLTs of the patterns with the ground truth pattern (leftmost one) reveals that this one is indeed the closest to the ground truth.

Fig. 10.D shows comparison of two GANs, WGAN [39] and WGAN-GP [91], over the MNIST dataset using the method above over single digits and the entire dataset. It shows that both models produce distributions that are

very close to the ground truth, but for almost all classes WGAN-GP shows better performance.

The geometry score does not use auxiliary networks and is not limited to visual data. However, since it only takes topological properties into account (which do not change if for example the entire dataset is shifted by 1) assessing the visual quality of samples may be difficult based only on this score. Due to this, authors propose to use this score in conjunction with other measures such as FID when dealing with natural images. .

21. **Reconstruction Error.** For many generative models, the reconstruction error on the training set is often explicitly optimized (*e.g.* Variational Autoencoders [8]). It is therefore natural to evaluate generative models using a reconstruction error measure (*e.g.* L_2 norm) computed on a test set. In the case of GANs, given a generator G and a set of test samples $X = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(m)}\}$ the reconstruction error of G on X is defined as:

$$\mathcal{L}_{rec}(G, X) = \frac{1}{m} \sum_{i=1}^m \min_{\mathbf{z}} \|G(\mathbf{z}) - \mathbf{x}^{(i)}\|^2. \quad (15)$$

Since it is not possible to directly infer the optimal \mathbf{z} from \mathbf{x} , Xiang and Li [48] used the following alternative method. Starting from an all-zero vector, they performed gradient descent on the latent code to find the one that minimizes the L_2 norm between the sample generated from the code and the target one. Since the code is optimized instead of being computed from a feed-forward network, the evaluation process is time-consuming. Thus, they avoided performing this evaluation at every training iteration when monitoring the training process, and only used a reduced number of samples and gradient descent steps. Only for the final trained model, they performed an extensive evaluation on a larger test set, with a larger number of steps.

22. **Image Quality Measures (SSIM, PSNR and Sharpness Difference).** Some researchers have proposed to use measures from the image quality assessment literature for training and evaluating GANs. They are explained next.

- (a) The single-scale SSIM measure [49] is a well-characterized perceptual similarity measure that aims to discount aspects of an image that are not important for human perception. It compares corresponding pixels and their neighborhoods in two images, denoted by x and y , using three quantities—luminance (I), contrast (C), and structure (S):

$$I(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \quad C(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad S(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}$$

The variables μ_x , μ_y , σ_x , and σ_y denote mean and standard deviations of pixel intensity in a local image patch centered at either x or y

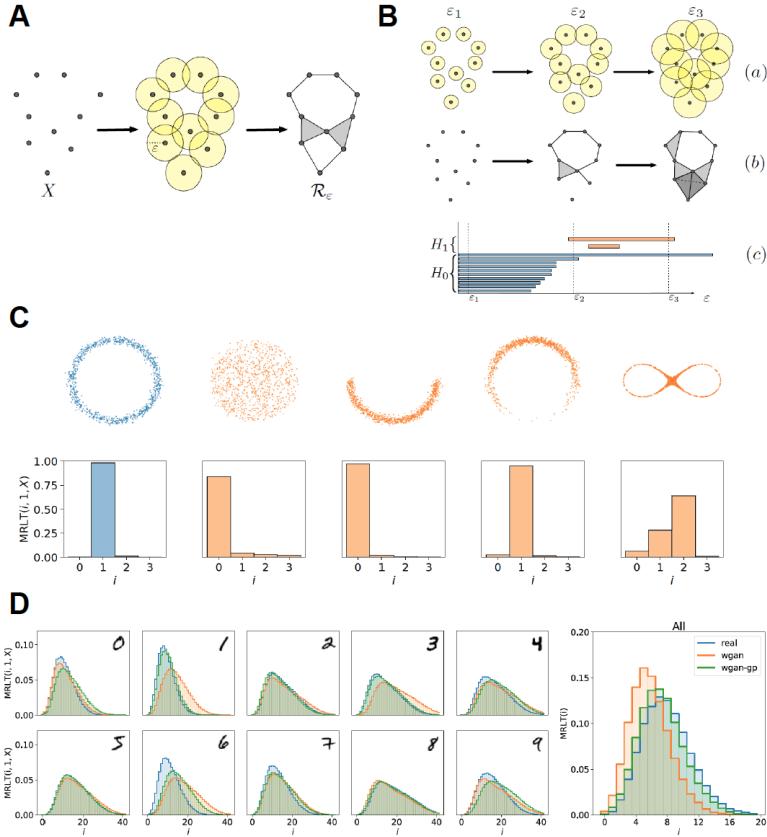


Figure 10: A) A simplicial complex constructed on a sample X . First, a fixed proximity parameter ϵ is chosen. Then, all balls of radius ϵ centered at each point are considered, and if for some subset of X of size $k + 1$ all the pairwise intersections of the corresponding balls are nonempty, then the k -dimensional simplex spanning this subset is added to the simplicial complex R_ϵ . B) Using different values for ϵ , different simplicial complexes are obtained (a). For $\epsilon = \epsilon_1$ the balls do not intersect and there are just 10 isolated components (b, [left]). For $\epsilon = \epsilon_2$ several components are merged and one loop is appeared (b, [middle]). The filled triangle corresponding to the triple pairwise intersection is topologically trivial and does not affect the topology (and similarly darker tetrahedron on the right). For $\epsilon = \epsilon_3$ all the components are merged into one and the same hole still exists (b, [right]). In the interval $[\epsilon_2, \epsilon_3]$ one smaller hole as in A is appeared and is quickly disappeared. This information is summarized in the persistence barcode (c). The number of connected components (holes) in the simplicial complex for some value ϵ_0 is given by the number of intervals in $H_0(H_1)$ intersecting the vertical line $\epsilon = \epsilon_0$. C) Mean Relative Living Times (MRLT) for various synthetic 2D datasets. The number of 1D holes is correctly identified in all the cases. Comparing MRLTs reveals that the second dataset from the left is closest to the ‘ground truth’ (noisy circle on the left). D) Comparison of MRLT of the MNIST dataset and of samples generated by WGAN and WGAN-GP trained on MNIST. MRLTs match almost perfectly, however, WGAN-GP shows slightly better performance on most of the classes. Figure compiled from [47].

(typically a square neighborhood of 5 pixels). The variable σ_{xy} denotes the sample correlation coefficient between corresponding pixels in the patches centered at x and y . The constants C_1 , C_2 , and C_3 are small values added for numerical stability. The three quantities are combined to form the SSIM score:

$$\text{SSIM}(x, y) = I(x, y)^\alpha C(x, y)^\beta S(x, y)^\gamma$$

SSIM assumes a fixed image sampling density and viewing distance. A variant of SSIM operates at multiple scales. The input images x and y are iteratively downsampled by a factor of 2 with a low-pass filter, with scale j denoting the original images downsampled by a factor of 2^{j-1} . The contrast $C(x, y)$ and structure $S(x, y)$ components are applied to all scales. The luminance component is applied only to the coarsest scale, denoted M . Further, contrast and structure components can be weighted at each scale. The final measure is:

$$\text{MS-SSIM}(x, y) = I_M(x, y)^{\alpha_M} \prod_{j=1}^M C_j(x, y)^{\beta_j} S_j(x, y)^{\gamma_j}$$

MS-SSIM ranges between 0 (low similarity) and 1 (high similarity). Snell *et al.* [50] defined a loss function for training GANs which is the sum of structural-similarity scores over all image pixels,

$$\mathcal{L}(X, Y) = - \sum_i \text{MS-SSIM}(X_i, Y_i),$$

where X and Y are the original and reconstructed images, and i is an index over image pixels. This loss function has a simple analytical derivative [92] which allows performing gradient descent. See Fig. 17 for more details.

- (b) PSNR measures the peak signal-to-noise ratio between two monochrome images I and K to assess the quality of a generated image compared to its corresponding real image (*e.g.* for evaluating conditional GANs [93]). The higher the PSNR (in db), the better quality of the generated image. It is computed as:

$$PSNR(I, K) = 10 \log_{10} \left(\frac{MAX_I^2}{MSE} \right) \quad (16)$$

$$= 20 \log_{10} (MAX_I) - 20 \log_{10} (MSE_{I,K}) \quad (17)$$

where

$$MSE_{I,K} = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} (I(m, n) - K(m, n))^2 \quad (18)$$

and, MAX_I is the maximum possible pixel value of the image (*e.g.* 255 for an 8 bit representation). This score can be used when a reference image is available for example in training conditional GANs using paired data (*e.g.* [15, 93]).

- (c) Sharpness Difference (SD) measures the loss of sharpness during image generation. It is compute as:

$$SD(I, K) = 10 \log_{10} \left(\frac{MAX_I^2}{GRADS_{I,K}} \right), \quad (19)$$

where

$$GRADS_{I,K} = \frac{1}{N} \sum_i \sum_j |(\nabla_i I + \nabla_j I) - (\nabla_i K + \nabla_j K)|, \quad (20)$$

and

$$\nabla_i I = |I_{i,j} - I_{i-1,j}|, \quad \nabla_j I = |I_{i,j} - I_{i,j-1}|. \quad (21)$$

Odena et al. [2] used ⁹ MS-SSIM to evaluate the diversity of generated images. The intuition is that image pairs with higher MS-SSIM seem more similar than pairs with lower MS-SSIM. They measured the MS-SSIM scores of 100 randomly chosen pairs of images within a given class. The higher (lower) diversity within a class, the lower (the higher) mean MS-SSIM score (See Fig. 11.A). Training images from the ImageNet training data contain a variety of mean MS-SSIM scores across the classes indicating the variability of image diversity in ImageNet classes. Fig. 11.B plots the mean MS-SSIM values for image samples versus training data for each class (after training was completed). It shows that 847 classes, out of 1000, have mean sample MS-SSIM scores below that of the maximum MS-SSIM for the training data. To identify whether the generator in AC-GAN [2] collapses during training, Odena et al. tracked the mean MS-SSIM score for all 1000 ImageNet classes (Fig. 11.C). Fig. 11.D shows the joint distribution of Inception accuracies versus MS-SSIM across all 1000 classes. It shows that Inception score and MS-SSIM are anti-correlated ($r^2 = -0.16$).

Juefei-Xu *et al.* [51] used the SSIM and PSNR measures to evaluate GANs in image completion tasks. The advantage here is that having 1-vs-1 comparison between the ground-truth and the completed image allows very straightforward visual examination of the GAN quality. It also allows head-to-head comparison between various GANs. In addition to the above mentioned image quality measures, some other measures such as Universal Quality Index (UQI) [94] and Visual Information Fidelity (VIF) [95] have also been adopted for assessing the quality of synthesized images. It has been reported that MS-SSIM finds large-scale mode collapses reliably but fails to diagnose smaller effects such as loss of variation in colors or textures. Its drawback is that it does not directly assess image quality in terms of similarity to the training set [2].

23. **Low-level Image Statistics.** Natural scenes make only a tiny fraction of the space of all possible images and have certain characteristics (*e.g.* [96,

⁹or ‘abused’ since the original MS-SSIM measure is intended to measure similarity of an image with respect to a reference image.

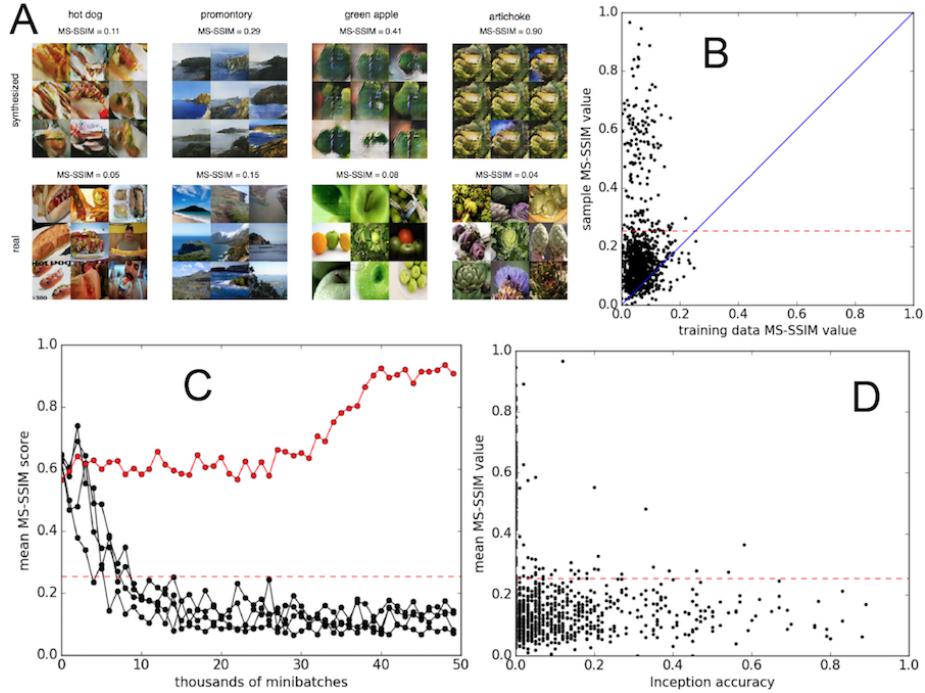


Figure 11: MS-SSIM score used for measuring image diversity. A) MS-SSIM scores for samples generated by the AC-GAN [2] model (top row) and training samples (bottom row). B) The mean MS-SSIM scores between pairs of images within a given class of the ImageNet dataset versus AC-GAN samples. The horizontal red line marks the maximum MS-SSIM across all ImageNet classes (over training data). Each data point represents the mean MS-SSIM value for samples from one class. C) Intra-class MS-SSIM for five ImageNet classes throughout a training run. Here, decreasing trend means successful training (black lines) whereas increasing trend indicates that the generator ‘collapses’ (red line). D) Comparison of Inception score vs. MS-SSIM for all 1000 ImageNet classes ($r^2 = -0.16$). AC-GAN samples do not achieve variability at the expense of discriminability. Figure compiled from [2].

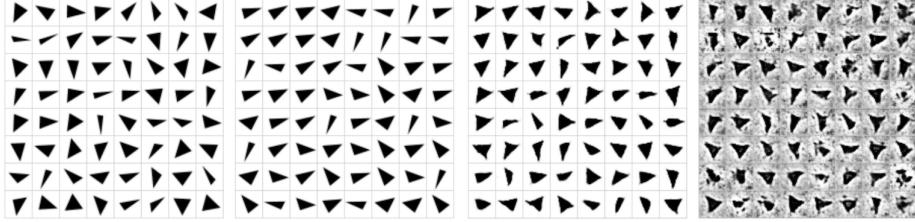
[\[97, 98, 99\]](#)). It has been shown that statistics of natural images remain the same when the images are scaled (*i.e.* scale invariance) [\[100, 101\]](#). The average power spectrum magnitude A over natural images has the form $A(f) = 1/f^{-\alpha}$, $\alpha \approx 2$ [\[102, 103, 104, 105\]](#). Another important property of natural image statistics is the non-Gaussianity [\[100, 101, 106\]](#). This means that marginal distribution of almost any zero mean linear filter response on virtually any dataset of images is sharply peaked at zero, with heavy tails and high kurtosis (greater than 3) [\[107\]](#). Recent studies have shown that the contrast statistics of the majority of natural images follows a Weibull distribution [\[108\]](#).

Zeng *et al.* [\[52\]](#) proposed to evaluate generative models in terms of low-level statistics of their generated images with respect to natural scenes. They considered four statistics including 1) the mean power spectrum, 2) the number of connected components in a given image area, 3) the distribution of random filter responses, and 4) the contrast distribution. Their results show that although generated images by DCGAN [\[1\]](#), WGAN [\[39\]](#), and VAE [\[19\]](#) resemble natural scenes in terms of low level statistics, there are still significant differences. For example, generated images do not have scale invariant mean power spectrum magnitude, which indicates existence of extra structures in these images caused by deconvolution operations.

Low-level image statistics can be used for regularizing GANs to optimize the discriminator to inspect whether the generator’s output matches expected statistics of the real samples (*a.k.a* feature matching [\[3\]](#)) using the loss function: $\|\mathbb{E}_{\mathbf{x} \sim P_r} f(\mathbf{x}) - \mathbb{E}_{\mathbf{z} \sim P_z(\mathbf{z})} f(G(\mathbf{z}))\|_2^2$, where $f(\cdot)$ represents the statistics of features. Karras *et al.* [\[53\]](#) investigated the multi scale statistical similarities between distributions of local image patches drawn from the Laplacian pyramid [\[109\]](#) representations of generated and real images. They used the Wasserstein distance to compare the distributions of patches¹⁰. The multi-scale pyramid allows a detailed comparison of statistics. The distance between the patch sets extracted from the lowest resolution indicates similarity in large-scale image structures, while the finest-level patches encode information about pixel-level attributes such as sharpness of edges and noise.

24. **Precision, Recall and F_1 Score.** Lucic *et al.* [\[23\]](#) proposed to compute precision, recall and F_1 score to quantify the degree of overfitting in GANs. Intuitively precision measures the quality of the generated samples, whereas recall measures the proportion of the reference distribution covered by the learned distribution. They argue that IS only captures precision as it does not penalize a model for not producing all modes of the data distribution. Rather, it only penalizes the model for not producing all classes. FID score, on the other hand, captures both precision and recall.
To approximate these scores for a model, Lucic *et al.* proposed to use toy datasets for which the data manifold is known and distances of generated

¹⁰This measure is known as the sliced Wasserstein distance (SWD)



(a) High precision, high recall (b) High precision, low recall (c) Low precision, high recall (d) Low precision, low recall

Figure 12: Samples from a model trained on gray-scale triangles. These triangles belong to a low-dimensional manifold embedded in $\mathbb{R}^{d \times d}$. A good generative model should be able to capture the factors of variation in this manifold (*e.g.* rotation, translation, minimum angle size). a) high recall and precision, b) high precision, but low recall (lacking in diversity), c) low precision, but **high recall** (can decently reproduce triangles, but fails to capture convexity), and d) low precision and low recall. Figure from [23].

samples to the manifold can be computed. An example of such dataset is the manifold of convex shapes (See Fig. 12). To compute these scores, first the latent representation \mathbf{z} of each test sample \mathbf{x} is estimated, through gradient descent, by inverting the generator G . Precision is defined as the fraction of the generated samples whose distance to the manifold is below a certain threshold. Recall, on the other hand, is given by the fraction of test samples whose L_2 distance to $G(\mathbf{z})$ is below the threshold. If the samples from the model distribution P_g are (on average) close to the manifold (see [23] for details), its precision is high. Similarly, high recall implies that the generator can recover (*i.e.* generate something close to) any sample from the manifold, thus capturing most of the manifold.

The major drawback of these scores is that they are impractical for real images where the data manifold is unknown, and their use is limited to evaluations on synthetic data. In a recent effort, Sajjadi et al. [72] introduced a novel definition of precision and recall to address this limitation.

2.3. Qualitative Measures

Visual examination of samples by human raters is one of the common and most intuitive ways to evaluate GANs (*e.g.* [110, 3, 111]). While it greatly helps inspect and tune models, it suffers from the following drawbacks. First, evaluating the quality of generated images with human vision is expensive and cumbersome, biased (*e.g.* depends on the structure and pay of the task, community reputation of the experimenter, etc in crowd sourcing setups [112]) difficult to reproduce, and does not fully reflect the capacity of models. Second, human inspectors may have high variance which makes it necessary to average over a large number of subjects. Third, an evaluation based on samples could be biased towards models that overfit and therefore a poor indicator of a good density model in a log-likelihood sense [22]. For instance, it fails to tell whether a model drops modes. In fact, mode dropping generally helps visual sample quality

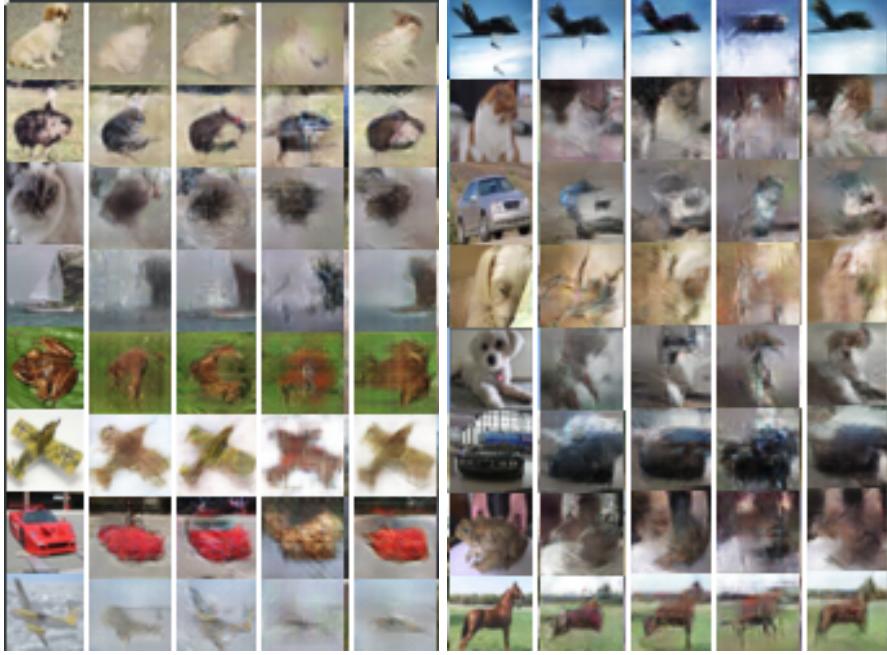


Figure 13: Generated samples nearest to real images from CIFAR-10. In each of the two panels, the first column shows real images, followed by the nearest image generated by DCGAN [1], ALI [114], Unrolled GAN [115], and VEEGAN [58]. Figure compiled from [58].

as the model can choose to focus on only few common modes that correspond to typical samples.

In what follows, I discuss the ways that have been followed in the literature to qualitatively inspect the quality of generated images by a model and explore its learned latent space.

1. **Nearest Neighbors.** To detect overfitting, traditionally some samples are shown next to their nearest neighbors in the training set (*e.g.* Fig. 13). There are, however, two concerns regarding this manner of evaluation:
 - (a) Nearest neighbors are typically determined based on the Euclidean distance which is very sensitive to minor perceptual perturbations. This is a well known phenomenon in the psychophysics literature (See Wang and Bovik [113]). It is trivial to generate samples that are visually almost identical to a training image, but have large Euclidean distances with it [22]. See Fig. 14.A for some examples.
 - (b) A model that stores (transformed) training images (*i.e.* memory GAN) can trivially pass the nearest-neighbor overfitting test [22]. This problem can be alleviated by choosing nearest neighbors based on perceptual measures, and by showing more than one nearest neighbor.

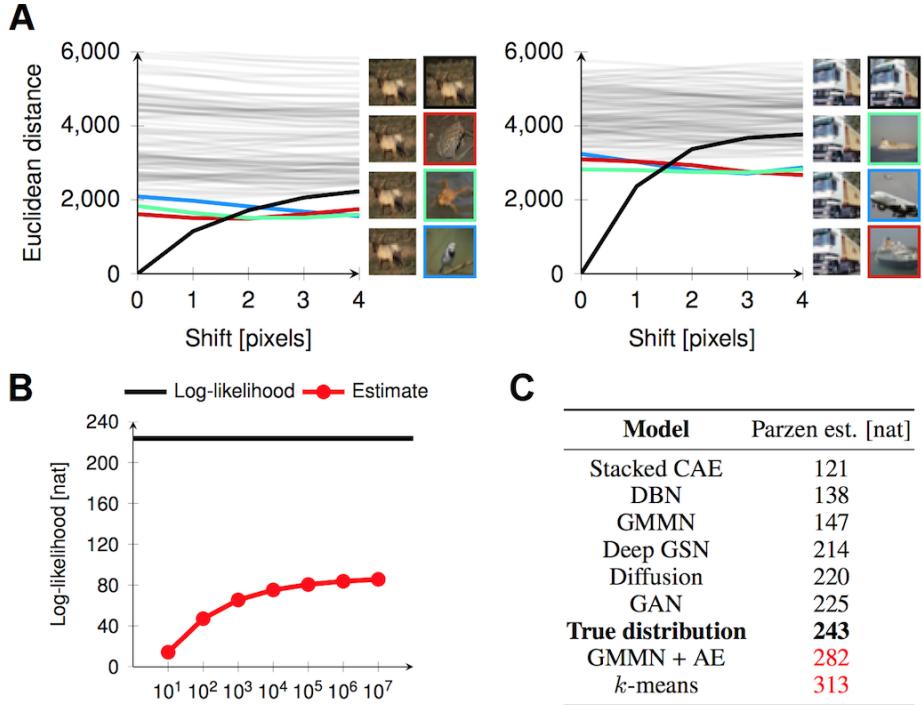


Figure 14: A) Small changes to an image can lead to large changes in Euclidean distance affecting the choice of the nearest neighbor. The left column shows a query image shifted 1 to 4 pixels (top to bottom). The right column shows the corresponding nearest neighbor from the training set. The gray lines indicate Euclidean distance of the query image to 100 randomly picked images from the training set. B) Parzen window estimates for a Gaussian evaluated on 6 by 6 pixel image patches from the CIFAR-10 dataset. Even for small patches and a very large number of samples, the Parzen window estimate is far from the true log-likelihood. C) Using Parzen window estimates to evaluate various models trained on MNIST, samples from the true distribution perform worse than samples from a simple model trained with *k*-means. Figure compiled from [22].

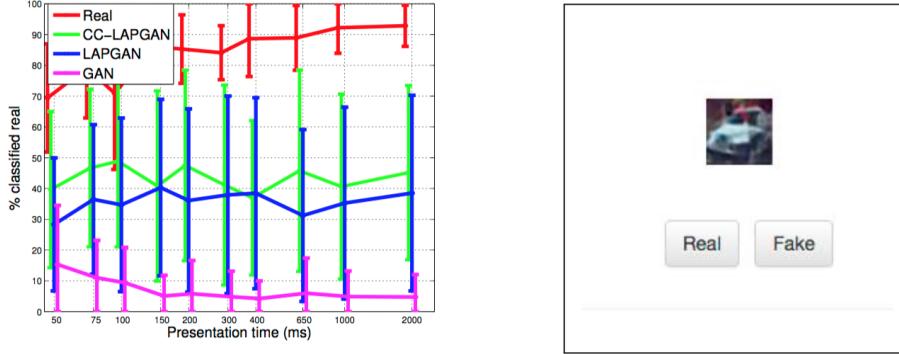


Figure 15: Left: Human evaluation of real CIFAR10 images (red) and samples from Goodfellow *et al.* [18] (magenta), and LAPGAN [110] and a class conditional LAPGAN (green). Around 40% of the samples generated by the class conditional LAPGAN model are realistic enough to fool a human into thinking they are real images. This compares with $\leq 10\%$ of images from the standard GAN model, but is still a lot lower than the $> 90\%$ rate for real images. Right: The user-interface presented to the subjects. Figure from [110].

2. Rapid Scene Categorization. These measures are inspired by prior studies who have shown that humans are capable of reporting certain characteristics of scenes in a short glance (*e.g.* scene category, visual layout [116, 117]). To obtain a quantitative measure of quality of samples, Denton *et al.* [110] asked volunteers to distinguish their generated samples from real images. The subjects were presented with the user interface shown in Fig. 15(right) and were asked to click the appropriate button to indicate if they believed the image was real or generated. They varied the viewing time from 50ms to 2000ms (11 durations). Fig. 15(left) shows the results over samples generated by three GAN models. They concluded that their model was better than the original GAN [18] since it did better in fooling the subjects (lower bound here is 0% and upper bound is 100%). See also Fig. 16 for another example of fake vs. real experiment but without time constraints (conducted by Salimans *et al.* [3]).

This ‘‘Turing-like’’ test is very intuitive and seems inevitable to ultimately answer the question of whether generative models are as good as the nature in generating images. However, there are several concerns in conducting such a test in practice (especially when dealing with models that are far from perfect; See Fig. 15(left)). Aside from experimental conditions which are hard to control in crowd-sourced platforms (*e.g.* presentation time, screen size, subject’s distance to the screen, subjects’ motivations, age, mood, feedback, etc) and high cost, these tests fall short in evaluating models in terms of diversity of generated samples and may be biased towards models that overfit to training data.

3. Rating and Preference Judgment. These types of experiments ask subjects to rate models in terms of the fidelity of their generated images. For example, Snell *et al.*, [118] studied whether observers prefer

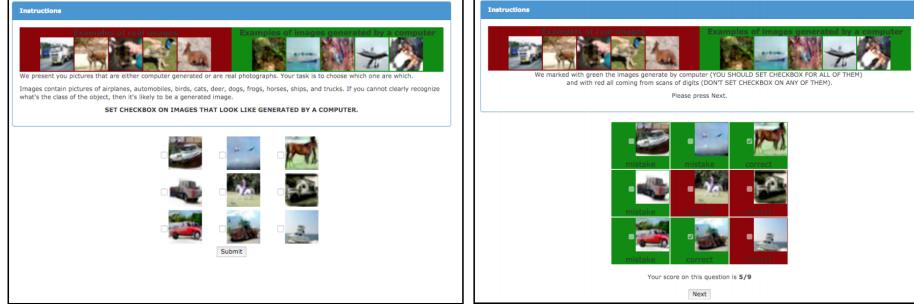


Figure 16: Web interface given to annotators in the experiments conducted by Salimans *et al.* [3]. Annotators are asked to distinguish computer generated images from real ones (left) and are provided with feedback (right). Figure compiled from [3].

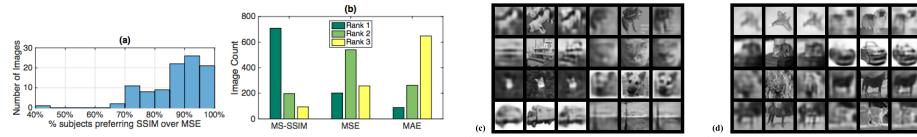


Figure 17: An example of a user judgment study by Snell *et al.* [118]. Left) Human judgments of generated images (a) Fully connected network: Proportion of participants preferring SSIM to MSE for each of 100 image triplets. (b) Deterministic conv. network: Distribution of image quality ranking for MS-SSIM, MSE, and MAE for 1000 images from the STL-10 hold-out set. Right) Image triplets consisting of—from left to right—the MSE reconstruction, the original image, and the SSIM reconstruction. Image triplets are ordered, from top to bottom and left to right, by the percentage of participants preferring SSIM. (c) Eight images for which participants strongly preferred SSIM over MSE. (d) Eight images for which the smallest proportion of participants preferred SSIM. Figure compiled from [118].

reconstructions produced by perceptually-optimized networks or by the pixelwise-loss optimized networks. Participants were shown image triplets with the original (reference) image in the center and the SSIM- and MSE-optimized reconstructions on either side with the locations counterbalanced. Participants were instructed to select which of the two reconstructed images they preferred (See Fig. 17). Similar approaches have been followed in [54, 55, 56, 57, 84, 119, 120, 121, 122]. Often the first few trials in these experiments are spared for practice.

4. **Evaluating Mode Drop and Mode Collapse.** GANs have been repeatedly criticized for failing to model the entire data distribution, while being able to generate realistically looking images. Mode collapse, *a.k.a* the Helvetica scenario, is the phenomenon when the generator learns to map several different input z vectors to the same output (possibly due to low model capacity or inadequate optimization [27]). It causes lack of diversity in the generated samples as the generator assigns low probability mass to significant subsets of the data distribution’s support. Mode drop occurs when some hard-to-represent modes of P_r are simply “ignored” by P_g . This is different than mode collapse where several modes of P_r are “averaged”

by P_g into a single mode, possibly located at a midpoint. An ideal GAN evaluation measure should be sensitive to these two phenomena.

Detecting mode collapse in GANs trained on large scale image datasets is very challenging¹¹. However, it can be accurately measured on synthetic datasets where the true distribution and its modes are known (*e.g.* Gaussian mixtures). Srivastava *et al.* [58] proposed a measure to quantify mode collapse behavior as follows:

- (a) First, some points are sampled from the generator. A sample is counted as *high quality*, if it is within a certain distance of its nearest mode center (*e.g.* 3σ over a 2D dataset, or 10σ over a 1200D dataset).
- (b) Then, the *number of modes captured* is the number of mixture components whose mean is nearest to at least one high quality sample. Accordingly, a mode is considered lost if there is no sample in the generated test data within a certain standard deviations from the center of that mode. This is illustrated in Fig. 19.

Santurkar et al. [42], to investigate mode distribution/collapse over natural datasets, propose to train GANs over a well-balanced dataset (*i.e.* a dataset that contains equal number of samples from each class) and then test whether generated data also generates a well-balanced dataset. Steps are as follows:

- (a) Train the GAN unconditionally (without class labels) on the chosen balanced multi-class dataset D.
- (b) Train a multi-class classifier on the same dataset D (to be used as an annotator).
- (c) Generate a synthetic dataset by sampling N images from the GAN. Then use the classifier trained in Step 2 above to obtain labels for this synthetic dataset.

An example is shown in Fig. 18. It reveals that GANs often exhibit mode collapse.

The reverse KL divergence over the modes has been used in [59] to measure the quality of mode collapse as follows. Each generated sample is assigned to its closest mode. This induces an empirical, discrete distribution with an alphabet size equal to the number of observed modes in the generated samples. A similar induced discrete distribution is computed from the real data samples. The reverse KL divergence between the induced distribution from generated samples and the induced distribution from the real samples is used as a measure.

The shortcoming of the described measures is that they only work for datasets with known modes (*e.g.* synthetic or labeled datasets). Overall, it is hard to quantitatively measure mode collapse and mode drop since they are poorly understood. Further, finding nearest neighbors and nearest mode center is non-trivial in high dimensional spaces is non-trivial. Active research is ongoing in this direction.

¹¹See [58, 26] for analysis of mode drop and mode collapse over real datasets.

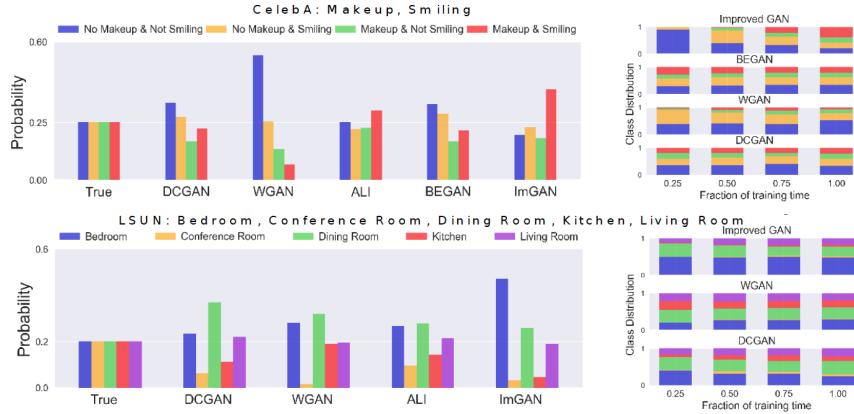


Figure 18: Illustration of mode collapse in GANs trained on select subsets of CelebA and LSUN datasets using the technique in [42]. Left panel shows the relative distribution of modes in samples drawn from the GANs, and compares is to the true data distribution (leftmost plots). Right panel shows the evolution of class distributions in different GANs over the course of training. It can be seen that these GANs introduce covariate shift through mode collapse. Figure compiled from [42].

5. Investigating and Visualizing the Internals of Networks. Other ways of evaluating generative models are studying how and what they learn, exploring their internal dynamics, and understanding the landscape of their latent spaces. While this is a broad topic and many papers fall under it, here I bring few examples to give the reader some insights.

- (a) *Disentangled representations.* “Disentanglement” regards the alignment of “semantic” visual concepts to axes in the latent space. Some tests can check the existence of semantically meaningful directions in the latent space, meaning that varying the seed along those directions leads to predictable changes (*e.g.* changes in facial hair, or pose). Some others (*e.g.* [60, 61, 62, 123]) assess the quality of internal representations by checking whether they satisfy certain properties, such as being “disentangled”. A measure of disentanglement proposed in [61] checks whether the latent space captures the true factors of variation in a simulated dataset where parameters are known by construction (*e.g.* using a graphics engine). Radford *et al.* [1] investigated their trained generators and discriminators in a variety of ways. They proposed that walking on the learned manifold can tell us about signs of memorization (if there are sharp transitions) and about the way in which the space is hierarchically collapsed. If walking in this latent space results in semantic changes to the image generations (such as objects being added and removed), one can reason that the model has learned relevant and interesting representations. They also showed interesting results of performing vector arithmetic on the z vectors of sets of exemplar samples for visual concepts (*e.g.* smiling woman



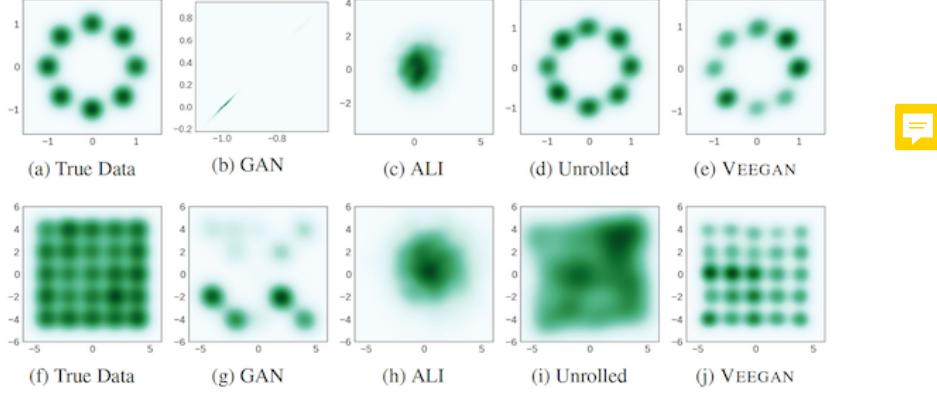


Figure 19: Density plots of the true data and generator distributions from different GAN methods trained on mixtures of Gaussians arranged in a ring (top) or a grid (bottom). Figure from [58].

- neutral woman + neutral man = smiling man; using z 's averaged over several samples).
- (b) *Space continuity.* Related to above, the goal here is to study the level of detail a model is capable of extracting. For example, given two random seed vectors z_1 and z_2 that generated two realistic images, we can check the images produced using seeds lying on the line joining z_1 and z_2 . If such “interpolated” images are reasonable and visually appealing, then this may be taken as a sign that a model can produce novel images rather than simply memorizing them (*e.g.* [124]; See Fig. 20). Some other examples include [120, 125]. White [126] suggests that replacing linear interpolation with spherical linear interpolation prevents diverging from a model’s prior distribution and produces sharper samples. Vedantam *et al.* [127] studied “visually grounded semantic imagination” and proposed several ways to evaluate their models in terms of the quality of the learned semantic latent space.
- (c) *Visualizing the discriminator features.* Motivated by previous studies on investigating the representations and features learned by convolutional neural networks trained for scene classification (*e.g.* [63, 64, 128]), some works have attempted to visualize the internal parts of generators and discriminators in GANs. For example, Radford *et al.* [1] showed that DCGAN trained on a large image dataset can also learn a hierarchy of interesting features. Using guided backpropagation [129], they showed that the features learned by the discriminator fire on typical parts of a bedroom, such as beds and windows (See Fig. 5 in [1]). The t-SNE method [130] has also been frequently used to project the learned latent spaces in 2D.

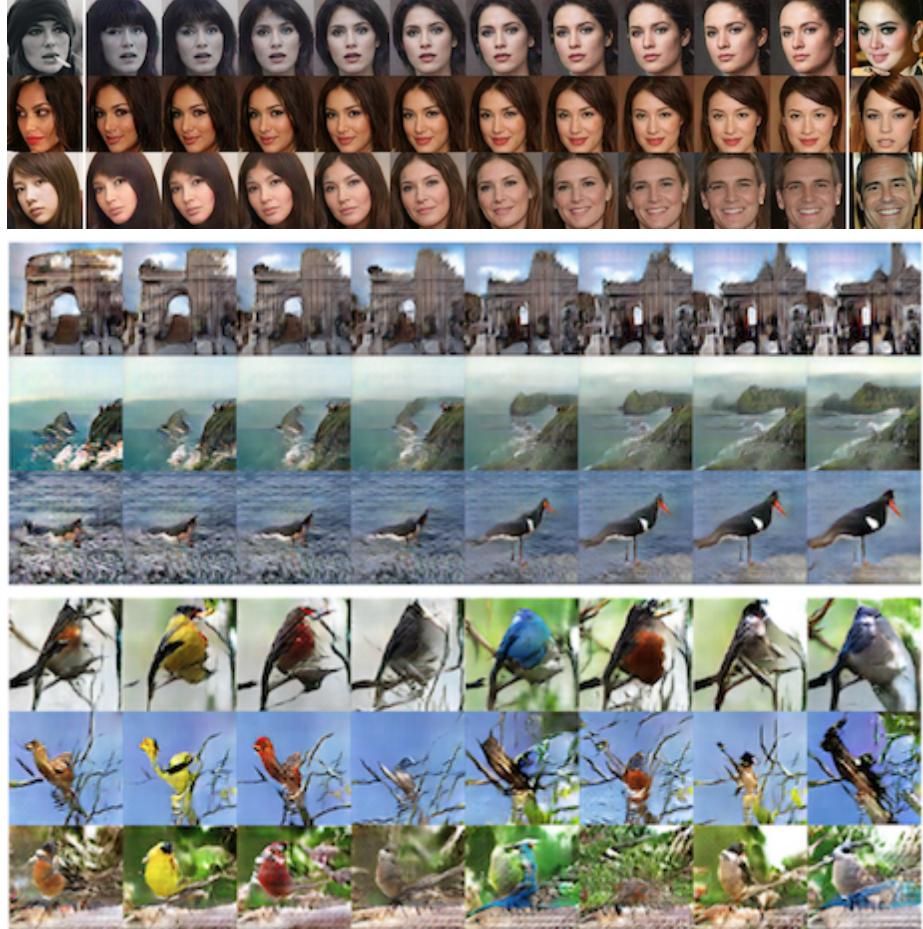


Figure 20: Top: Interpolations on z_r between real images at 128×128 resolution (from BEGAN [124]). These images were not part of the training data. The first and last columns contain the real images to be represented and interpolated. The images immediately next to them are their corresponding approximations while the images in between are the results of linear interpolation in z_r . Middle: Latent space interpolations for three ImageNet classes. Left-most and right-columns show three pairs of image samples - each pair from a distinct class. Intermediate columns highlight linear interpolations in the latent space between these three pairs of images (From [2]). Bottom: Class-independent information contains global structure about the synthesized image. Each column is a distinct bird class while each row corresponds to a fixed latent code z (From [2]).

3. Discussion

3.1. Other Evaluation Measures

In addition to measures discussed above, there exist some other non-trivial or task-specific ways to evaluate GANs. Vedantam *et al.* [127] proposed a model for visually grounded imagination to create images of novel semantic concepts. To evaluate the quality of the generated images, they proposed three measures including *a) correctness*: fraction of attributes for each generated image that match those specified in the concept’s description, *b) coverage*: diversity of values for the unspecified or missing attributes, measured as the difference between the empirical distributions of attribute values in the generated set and the true distribution for this attribute induced by the training set, and *c) compositionality*: correctness of generated images in response to test concepts that differ in at least one attribute from the training concepts. To measure diversity of generated samples, Zhu *et al.* [131] randomly sampled from their model and computed the average pair-wise distance in a deep feature space using cosine distance and compared it with the same measure calculated from ground truth real images. This is akin to the image retrieval performance measure described above. Im *et al.* [132] proposed to evaluate GANs by exploring the divergence and distance measures that were used during training GANs. They showed that rankings produced by four measures including 1) Jensen-Shannon Divergence, 2) Constrained Pearson χ^2 , 3) Maximum Mean Discrepancy, and 4) Wasserstein Distance, are consistent and robust across measures.

3.2. Sample and Computational Efficiencies

Here, I provide more details on two items in the list of desired properties of GAN evaluation measures. They will be used in the next subsection for assessing the measures. Huang *et al.* [26] argue that a practical GAN evaluation measure should be computed using a reasonable number of samples and within an affordable computation cost. This is particularly important during monitoring the training process of models. They proposed the following ways to assess evaluation measures:

1. *Sample efficiency*: It regards the number of samples needed for a measure to discriminate a set of generated samples S_g from a set of real samples S'_r . To do this, a reference set S_r is uniformly sampled from the real training data (but disjoint with S'_r). All three sets have the same size (*i.e.* $|S_r| = |S'_r| = |S_g| = n$). An ideal measure ρ is expected to correctly score $\rho(S_r, S'_r)$ lower than $\rho(S_r, S_g)$ with a relatively small n . In other words, the number of samples n needed for a measure to distinguish S'_r and S_g can be viewed as its sample complexity.
2. *Computational efficiency*: Fast computation of the empirical measure is of practical concern as it helps researchers monitor the training process and diagnose problems early on (*e.g.* for early stopping). This can be measured in terms of seconds per number of evaluated samples.

Measure	Desiderata							
	Discriminability	Detecting Overfitting	Disentangled Latent Spaces	Well-defined Bounds	Perceptual Judgments	Sensitivity to Distortions	Comp. & Sample Efficiency	
1. Average Log- likelihood [18, 22]	low	low	-	[-∞, ∞]	low	low	low	
2. Coverage Metric [33]	low	low	-	[0, 1]	low	low	-	
3. Inception Score (IS) [3]	high	moderate	-	[1, ∞]	high	moderate	high	
4. Modified Inception Score (m-IS) [34]	high	moderate	-	[1, ∞]	high	moderate	high	
5. Mode Score (MS) [35]	high	moderate	-	[0, ∞]	high	moderate	high	
6. AM Score [36]	high	moderate	-	[0, ∞]	high	moderate	high	
7. Fréchet Inception Distance (FID) [37]	high	moderate	-	[0, ∞]	high	high	high	
8. Maximum Mean Discrepancy (MMD) [38]	high	low	-	[0, ∞]	-	-	-	
9. The Wasserstein Critic [39]	high	moderate	-	[0, ∞]	-	-	low	
10. Birthday Paradox Test [27]	low	high	-	[1, ∞]	low	low	-	
11. Classifier Two Sample Test (C2ST) [40]	high	low	-	[0, 1]	-	-	-	
12. Classification Performance [1, 15]	high	low	-	[0, 1]	low	-	-	
13. Boundary Distortion [42]	low	low	-	[0, 1]	-	-	-	
14. NDB [43]	low	high	-	[0, ∞]	-	low	-	
15. Image Retrieval Performance [44]	moderate	low	-	*	low	-	-	
16. Generative Adversarial Metric (GAM) [31]	high	low	-	*	-	-	moderate	
17. Tournament Win Rate and Skill Rating [45]	high	high	high	-	*	-	low	
18. NRDS [32]	high	low	-	[0, 1]	-	-	poor	
19. Adversarial Accuracy & Divergence [46]	high	low	-	[0, 1], [0, ∞]	-	-	-	
20. Geometry Score [47]	low	low	-	[0, ∞]	-	low	low	
21. Reconstruction Error [48]	low	low	-	[0, ∞]	-	moderate	moderate	
22. Image Quality Measures [49, 50, 51]	low	moderate	-	*	high	high	high	
23. Low-level Image Statistics [52, 53]	low	low	-	*	low	low	-	
24. Precision, Recall and F_1 score [23]	low	high	✓	[0, 1]	-	-	-	

Table 2: Meta measure of GAN quantitative evaluation scores. Notice that the ratings are relative. “-” means unknown (hence warranting further research). “*” indicates that several bounds for several scores in that family measure are available. Also, notice that tighter bounds for some of the measures might be possible. It seems that most of the measures do not systematically evaluate disentanglement in the latent space.

3.3. What is the Best GAN Evaluation Measure?

To answer this question, let's first take a look at how well the measures perform with respect to the desired properties mentioned in Section 2.1. Results are shown in Table 2. I find that:

- (a) only two measures are designed to explicitly address overfitting,
- (b) the majority of the measures do not consider disentangled representations,
- (c) few measures have both lower and upper bounds,
- (d) the agreement between the measures and human perceptual judgments is less clear,
- (e) several highly regarded measures have high sample and computational efficiencies, and
- (f) the sensitivity of measures to image distortions is less explored.

A detailed discussion and comparison of GAN evaluation measures comes next.

As of yet, there is no consensus regarding the best score. Different scores assess various aspects of the image generation process, and it is unlikely that a single score can cover all aspects. Nevertheless, some measures seem more plausible than others (*e.g.* FID score). Detailed analyses by Theis *et al.* [22] showed that average likelihood is not a good measure. Parzen windows estimation of likelihood favors trivial models and is irrelevant to visual fidelity of samples. Further, it fails to approximate the true likelihood in high dimensional spaces or to rank models (Fig. 14). Similarly, the Wasserstein distance between generated samples and the training data is also intractable in high dimensions [53]. Two widely accepted scores, Inception Score and Fréchet Inception Distance, rely on pre-trained deep networks to represent and statistically compare original and generated samples. This brings along two significant drawbacks. First, the deep network is trained to be invariant to image transformations and artifacts making the evaluation method also insensitive to those distortions. Second, since the deep network is often trained on large scale natural scene datasets (*e.g.* ImageNet), applying them to other domains (*e.g.* faces, digits) is questionable. Some evaluation methods (*e.g.* MS-SSIM [2], Birthday Paradox Test) aim to assess the diversity of the generated samples, regardless of the data distribution. While being able to detect severe cases of mode collapse, these methods fall short in measuring how well a generator captures the true data distribution [53].

Quality measures such as nearest neighbor visualizations or rapid categorization tasks may favor models that overfit. Overall, it seems that the main challenge is to have a measure that evaluates both *diversity* and *visual fidelity* simultaneously. The former implies that all modes are covered while the latter implies that the generated samples should have high likelihood. Perhaps due to these challenges, Theis *et al.* [22] argued against evaluating models for task-independent image generation and proposed to evaluate GANs with respect to specific applications. For different applications then,

different measures might be more appropriate. For example, the likelihood is good for measuring compression methods [7] while psychophysics and user ratings are fit for evaluating image reconstruction and synthesis methods [8, 133]. Some measures are suitable for evaluating generic GANs (when input is a noise vector), while some others are suitable for evaluating conditional GANs (*e.g.* FCN score) where correspondences are available (*e.g.* generating an image corresponding to a segmentation map).

Despite having different formulations, several scores are based on similar concepts. C2ST, adversarial accuracy, and classification performance employ classifiers to determine how separable generated images are from real images (on a validation dataset). FID, Wasserstein and MMD measures compute the distance between two distributions. Inception score and its variants including m-IS, Mode and AM scores use conditional and marginal distributions over generated data or real data to evaluate diversity and fidelity of samples. Average log likelihood and coverage metric estimate the probability distributions. Reconstruction error and some quality measures determine how dissimilar generated images are from their corresponding (or closest) images in the train set. Some measures use individual samples (*e.g.* IS) while others need pairs of samples (*e.g.* MMD). One important concern regarding many measures is that they are sensitive to the choice of the feature space (*e.g.* different CNNs) as well as the distance type (*e.g.* L_2 vs. L_1).

Fidelity, diversity and controllable sampling are the main aspects of a model that a measure should capture. A good score should have well defined bounds and also be sensitive to image distortions and transformations (See Fig. 21 and 4). One major problem with qualitative measures such as SSIM and PSNR is that they only tap visual fidelity and not diversity of samples. Humans are also often biased towards the visual quality of generated images and are less affected by the lack of image diversity. On the other hand, some quantitative measures mostly concentrate on evaluating diversity (*e.g.* Birthday Paradox Test) and discard fidelity. Ideally, a good measure should take both into account.

Fig. 22 shows a comparison of GAN evaluation measures in terms of sample and computational efficiency. While some measures are practical to compute for a small sample size (about 2000 images), some others (*e.g.* Wasserstein distance) do not scale to large sample sizes. Please see [26] for further details.

4. Summary and Future Work

In this work, I provided a critical review of the strengths and limitations of 24 quantitative and 5 qualitative measures that have been introduced so far for evaluating GANs. Seeking appropriate measures for this purpose continues to be an important open problem, not only for fair model comparison but also for understanding, improving, and developing generative models. Lack of a

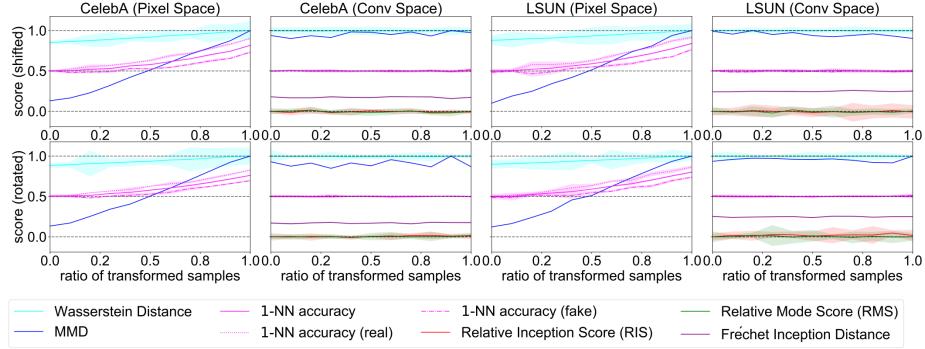


Figure 21: Robustness analysis of different GAN evaluation measures to small image transformations (rotations and translations). A good measure is expected to remain constant across all mixes of real and transformed real samples, since the transformations do not alter semantics of the image. Some measures are more susceptible to changes in the pixel space than the convolutional space. Figure from [26].

universal powerful measure can hinder the progress. In a recent benchmark study, Lucic *et al.* [23] found no empirical evidence in favor of GAN models who claimed superiority over the original GAN. In this regard, borrowing from other fields such as natural scene statistics and cognitive vision can be rewarding. For example, understanding how humans perceive symmetry [134, 135] or image clutter [136] in generated images versus natural scenes can give clues regarding the plausibility of the generated images.

Ultimately, I suggest the following directions for future research in this area:

1. creating a code repository of evaluation measures,
2. conducting detailed comparative empirical and analytical studies of available measures, and
3. benchmarking models under the same conditions (*e.g.* architectures, optimization, hyperparameters, computational budget) using more than one measure.

References

- [1] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv preprint arXiv:1511.06434.
- [2] A. Odena, C. Olah, J. Shlens, Conditional image synthesis with auxiliary classifier gans, arXiv preprint arXiv:1610.09585.
- [3] T. Salimans, W. Goodfellow, Ian gotand Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training gans, in: Advances in Neural Information Processing Systems, 2016, pp. 2234–2242.

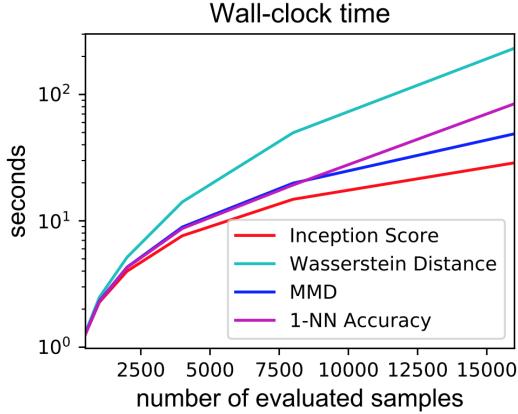


Figure 22: Measurement of wall-clock time for computing various measures as a function of the number of samples. As it shows, all measures are practical to compute for a sample of size 2000, but Wasserstein distance does not scale to large sample sizes. Figure from [26].

- [4] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks, *The Journal of Machine Learning Research* 17 (1) (2016) 2096–2030.
- [5] E. Tzeng, J. Hoffman, K. Saenko, T. Darrell, Adversarial discriminative domain adaptation, in: Computer Vision and Pattern Recognition (CVPR), Vol. 1, 2017, p. 4.
- [6] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis, arXiv preprint arXiv:1605.05396.
- [7] L. Theis, W. Shi, A. Cunningham, F. Huszar, Lossy image compression with compressive autoencoders, arXiv preprint arXiv:1703.00395.
- [8] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network, arXiv preprint.
- [9] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A. A. Efros, Context encoders: Feature learning by inpainting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2536–2544.
- [10] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, M. N. Do, Semantic image inpainting with deep generative models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5485–5493.

- [11] J. Pan, C. Canton, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, X. Giro-i Nieto, Salgan: Visual saliency prediction with generative adversarial networks, arXiv preprint arXiv:1701.01081.
- [12] H. Zhang, V. Sindagi, V. M. Patel, Image de-raining using a conditional generative adversarial network, arXiv preprint arXiv:1701.05957.
- [13] L. A. Gatys, A. S. Ecker, M. Bethge, Image style transfer using convolutional neural networks, in: Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on, IEEE, 2016, pp. 2414–2423.
- [14] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: European Conference on Computer Vision, Springer, 2016, pp. 694–711.
- [15] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, arXiv preprint.
- [16] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, arXiv preprint arXiv:1703.10593.
- [17] C. Vondrick, H. Pirsiavash, A. Torralba, Generating videos with scene dynamics, in: Advances In Neural Information Processing Systems, 2016, pp. 613–621.
- [18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680.
- [19] D. P. Kingma, M. Welling, Auto-encoding variational bayesge, arXiv preprint arXiv:1312.6114.
- [20] D. P. Kingma, S. Mohamed, D. J. Rezende, M. Welling, Semi-supervised learning with deep generative models, in: Advances in Neural Information Processing Systems, 2014, pp. 3581–3589.
- [21] A. van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al., Conditional image generation with pixelcnn decoders, in: Advances in Neural Information Processing Systems, 2016, pp. 4790–4798.
- [22] L. Theis, A. v. d. Oord, M. Bethge, A note on the evaluation of generative models, arXiv preprint arXiv:1511.01844.
- [23] M. Lucic, K. Kurach, M. Michalski, S. Gelly, O. Bousquet, Are gans created equal? a large-scale study, arXiv preprint arXiv:1711.10337.
- [24] K. Kurach, M. Lucic, X. Zhai, M. Michalski, S. Gelly, The gan landscape: Losses, architectures, regularization, and normalization, arXiv preprint arXiv:1807.04720.

- [25] K. Shmelkov, C. Schmid, K. Alahari, How good is my gan?, ECCV.
- [26] G. Huang, Y. Yuan, Q. Xu, C. Guo, Y. Sun, F. Wu, K. Weinberger, [An empirical study on evaluation metrics of generative adversarial networks](#), International Conference on Learning RepresentationsRejected.
URL <https://openreview.net/forum?id=Sy1f0e-R>
- [27] S. Arora, Y. Zhang, Do gans actually learn the distribution? an empirical study, arXiv preprint arXiv:1706.08224.
- [28] N. Chen, A. Klushyn, R. Kurle, X. Jiang, J. Bayer, P. van der Smagt, Metrics for deep generative models, arXiv preprint arXiv:1711.01204.
- [29] Y. Wu, Y. Burda, R. Salakhutdinov, R. Grosse, On the quantitative analysis of decoder-based generative models, arXiv preprint arXiv:1611.04273.
- [30] A. A. Alemi, B. Poole, I. Fischer, J. V. Dillon, R. A. Saurous, K. Murphy, An information-theoretic analysis of deep latent-variable models, arXiv preprint arXiv:1711.00464.
- [31] D. J. Im, C. D. Kim, H. Jiang, R. Memisevic, Generating images with recurrent adversarial networks, arXiv preprint arXiv:1602.05110.
- [32] Z. Zhang, Y. Song, H. Qi, Decoupled learning for conditional adversarial networks, arXiv preprint arXiv:1801.06790.
- [33] I. O. Tolstikhin, S. Gelly, O. Bousquet, C.-J. Simon-Gabriel, B. Schölkopf, Adagan: Boosting generative models, in: Advances in Neural Information Processing Systems, 2017, pp. 5430–5439.
- [34] S. Gurumurthy, R. K. Sarvadevabhatla, V. B. Radhakrishnan, Deligan: Generative adversarial networks for diverse and limited data, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 1, 2017.
- [35] T. Che, Y. Li, A. P. Jacob, Y. Bengio, W. Li, Mode regularized generative adversarial networks, arXiv preprint arXiv:1612.02136.
- [36] Z. Zhou, H. Cai, S. Rong, Y. Song, K. Ren, W. Zhang, J. Wang, Y. Yu, [Activation maximization generative adversarial nets](#), International Conference on Learning Representations.
URL <https://openreview.net/forum?id=HyyP33gAZ>
- [37] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, in: Advances in Neural Information Processing Systems, 2017, pp. 6629–6640.
- [38] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, A. Smola, A kernel two-sample test, Journal of Machine Learning Research 13 (Mar) (2012) 723–773.

- [39] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein gan, arXiv preprint arXiv:1701.07875.
- [40] E. L. Lehmann, J. P. Romano, Testing statistical hypotheses, Springer Science & Business Media, 2006.
- [41] Y. Ye, L. Wang, Y. Wu, Y. Chen, Y. Tian, Z. Liu, Z. Zhang, Gan quality index (gqi) by gan-induced classifier.
- [42] S. Santurkar, L. Schmidt, A. Madry, A classification-based study of covariate shift in gan distributions, in: International Conference on Machine Learning, 2018, pp. 4487–4496.
- [43] E. Richardson, Y. Weiss, On gans and gmms, arXiv preprint arXiv:1805.12462.
- [44] Y. Wang, L. Zhang, J. van de Weijer, Ensembles of generative adversarial networks, arXiv preprint arXiv:1612.00991.
- [45] C. Olsson, S. Bhupatiraju, T. Brown, A. Odena, I. Goodfellow, Skill rating for generative models, arXiv preprint arXiv:1808.04888.
- [46] J. Yang, A. Kannan, D. Batra, D. Parikh, Lr-gan: Layered recursive generative adversarial networks for image generation, arXiv preprint arXiv:1703.01560.
- [47] V. Khrulkov, I. Oseledets, Geometry score: A method for comparing generative adversarial networks, arXiv preprint arXiv:1802.02664.
- [48] S. Xiang, H. Li, On the effects of batch and weight normalization in generative adversarial networks, stat 1050 (2017) 22.
- [49] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE transactions on image processing 13 (4) (2004) 600–612.
- [50] K. Ridgeway, J. Snell, B. Roads, R. S. Zemel, M. C. Mozer, Learning to generate images with perceptual similarity metrics, arXiv preprint arXiv:1511.06409.
- [51] F. Juefei-Xu, V. N. Boddeti, M. Savvides, Gang of gans: Generative adversarial networks with maximum margin ranking, arXiv preprint arXiv:1704.04865.
- [52] Y. Zeng, H. Lu, A. Borji, Statistics of deep generated images, arXiv preprint arXiv:1708.02688.
- [53] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of gans for improved quality, stability, and variation, arXiv preprint arXiv:1710.10196.

- [54] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, S. Belongie, Stacked generative adversarial networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 2, 2017, p. 4.
- [55] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, D. Metaxas, Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, in: IEEE Int. Conf. Comput. Vision (ICCV), 2017, pp. 5907–5915.
- [56] C. Xiao, B. Li, J.-Y. Zhu, W. He, M. Liu, D. Song, Generating adversarial examples with adversarial networks, arXiv preprint arXiv:1801.02610.
- [57] Z. Yi, H. Zhang, P. Tan, M. Gong, Dualgan: Unsupervised dual learning for image-to-image translation, arXiv preprint.
- [58] A. Srivastava, L. Valkoz, C. Russell, M. U. Gutmann, C. Sutton, Veegan: Reducing mode collapse in gans using implicit variational learning, in: Advances in Neural Information Processing Systems, 2017, pp. 3310–3320.
- [59] Z. Lin, A. Khetan, G. Fanti, S. Oh, Pacgan: The power of two samples in generative adversarial networks, arXiv preprint arXiv:1712.04086.
- [60] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, P. Abbeel, Infogan: Interpretable representation learning by information maximizing generative adversarial nets, in: Advances in Neural Information Processing Systems, 2016, pp. 2172–2180.
- [61] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, A. Lerchner, beta-vae: Learning basic visual concepts with a constrained variational framework.
- [62] M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, Y. LeCun, Disentangling factors of variation in deep representation using adversarial training, in: Advances in Neural Information Processing Systems, 2016, pp. 5040–5048.
- [63] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European conference on computer vision, Springer, 2014, pp. 818–833.
- [64] D. Bau, B. Zhou, A. Khosla, A. Oliva, A. Torralba, Network dissection: Quantifying interpretability of deep visual representations, in: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, IEEE, 2017, pp. 3319–3327.
- [65] R. M. Neal, Annealed importance sampling, Statistics and computing 11 (2) (2001) 125–139.
- [66] F. Huszár, How (not) to train your generative model: Scheduled sampling, likelihood, adversary?, arXiv preprint arXiv:1511.05101.

- [67] W. Fedus, M. Rosca, B. Lakshminarayanan, A. M. Dai, S. Mohamed, I. Goodfellow, Many paths to equilibrium: Gans do not need to decrease adivergence at every step, arXiv preprint arXiv:1710.08446.
- [68] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
- [69] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, 2009, pp. 248–255.
- [70] S. Barratt, R. Sharma, A note on the inception score, arXiv preprint arXiv:1801.01973.
- [71] Z. Zhou, W. Zhang, J. Wang, Inception score, label smoothing, gradient vanishing and-log ($d(x)$) alternative, arXiv preprint arXiv:1708.01729.
- [72] M. S. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, S. Gelly, Assessing generative models via precision and recall, arXiv preprint arXiv:1806.00035.
- [73] S. Liu, Y. Wei, J. Lu, J. Zhou, An improved evaluation framework for generative adversarial networks, arXiv preprint arXiv:1803.07474.
- [74] R. Fortet, E. Mourier, Convergence de la répartition empirique vers la répartition théorique, in: Annales scientifiques de l’École Normale Supérieure, Vol. 70, Elsevier, 1953, pp. 267–285.
- [75] K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Schölkopf, et al., Kernel mean embedding of distributions: A review and beyond, Foundations and Trends® in Machine Learning 10 (1-2) (2017) 1–141.
- [76] D. J. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, A. Gretton, Generative models and model criticism via optimized maximum mean discrepancy, arXiv preprint arXiv:1611.04488.
- [77] W. Boulniphone, E. Belilovsky, M. B. Blaschko, I. Antonoglou, A. Gretton, A test of relative similarity for model selection in generative models, arXiv preprint arXiv:1511.04581.
- [78] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, B. Póczos, Mmd gan: Towards deeper understanding of moment matching network, in: Advances in Neural Information Processing Systems, 2017, pp. 2200–2210.
- [79] Y. Li, K. Swersky, R. Zemel, Generative moment matching networks, in: International Conference on Machine Learning, 2015, pp. 1718–1727.

- [80] G. K. Dziugaite, D. M. Roy, Z. Ghahramani, Training generative neural networks via maximum mean discrepancy optimization, arXiv preprint arXiv:1505.03906.
- [81] M. Bińkowski, D. J. Sutherland, M. Arbel, A. Gretton, Demystifying MMD gans, arXiv preprint arXiv:1801.01401.
- [82] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proceedings of International Conference on Computer Vision (ICCV), 2015.
- [83] D. Lopez-Paz, M. Oquab, Revisiting classifier two-sample tests, arXiv preprint arXiv:1610.06545.
- [84] R. Zhang, P. Isola, A. A. Efros, Colorful image colorization, in: European Conference on Computer Vision, Springer, 2016, pp. 649–666.
- [85] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.
- [86] T. Lesort, F. Bordes, J.-F. Goudou, D. Filliat, Evaluation of generative networks through their data augmentation capacity (2018).
URL <https://openreview.net/forum?id=HJ1HFIZAb>
- [87] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.
- [88] I. Durugkar, I. Gemp, S. Mahadevan, Generative multi-adversarial networks, arXiv preprint arXiv:1611.01673.
- [89] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A. Y. Ng, Reading digits in natural images with unsupervised feature learning, in: NIPS workshop on deep learning and unsupervised feature learning, Vol. 2011, 2011, p. 5.
- [90] M. E. Glickman, A comprehensive guide to chess ratings, American Chess Journal 3 (1995) 59–102.
- [91] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville, Improved training of wasserstein gans, arXiv preprint arXiv:1704.00028.
- [92] Z. Wang, E. P. Simoncelli, Maximum differentiation (mad) competition: A methodology for comparing computational models of perceptual quantities, Journal of Vision 8 (12) (2008) 8–8.
- [93] K. Regmi, A. Borji, Cross-view image synthesis using conditional gans, CVPR.

- [94] Z. Wang, A. Bovik, A universal image quality index, *IEEE Signal Processing Letters* 9 (3) (2002) 81–84. [doi:10.1109/97.995823](https://doi.org/10.1109/97.995823).
- [95] H. R. Sheikh, A. C. Bovik, Image information and visual quality, *IEEE Transactions on image processing* 15 (2) (2006) 430–444.
- [96] W. S. Geisler, Visual perception and the statistical properties of natural scenes, *Annu. Rev. Psychol.* 59 (2008) 167–192.
- [97] E. P. Simoncelli, B. A. Olshausen, Natural image statistics and neural representation, *Annual review of neuroscience* 24 (1) (2001) 1193–1216.
- [98] A. Torralba, A. Oliva, Statistics of natural image categories, *Network: computation in neural systems* 14 (3) (2003) 391–412.
- [99] D. L. Ruderman, The statistics of natural images, *Network: computation in neural systems* 5 (4) (1994) 517–548.
- [100] A. Srivastava, A. B. Lee, E. P. Simoncelli, S.-C. Zhu, On advances in statistical modeling of natural images, *Journal of mathematical imaging and vision* 18 (1) (2003) 17–33.
- [101] S.-C. Zhu, Statistical modeling and conceptualization of visual patterns, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25 (6) (2003) 691–712.
- [102] N. Deriugin, The power spectrum and the correlation function of the television signal, *Telecommunications* 1 (7) (1956) 1–12.
- [103] R. W. Cohen, I. Gorog, C. R. Carlson, Image descriptors for displays, Tech. rep., DTIC Document (1975).
- [104] G. Burton, I. R. Moorhead, Color and spatial structure in natural scenes, *Applied Optics* 26 (1) (1987) 157–170.
- [105] D. J. Field, Relations between the statistics of natural images and the response properties of cortical cells, *JOSA A* 4 (12) (1987) 2379–2394.
- [106] M. J. Wainwright, E. P. Simoncelli, Scale mixtures of gaussians and the statistics of natural images., in: *Nips*, 1999, pp. 855–861.
- [107] A. B. Lee, D. Mumford, J. Huang, Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model, *International Journal of Computer Vision* 41 (1) (2001) 35–59.
- [108] S. Ghebreab, S. Scholte, V. Lamme, A. Smeulders, A biologically plausible model for rapid natural scene identification, in: *Advances in Neural Information Processing Systems*, 2009, pp. 629–637.
- [109] P. J. Burt, E. H. Adelson, The laplacian pyramid as a compact image code, in: *Readings in Computer Vision*, Elsevier, 1987, pp. 671–679.

- [110] E. L. Denton, S. Chintala, R. Fergus, et al., Deep generative image models using a laplacian pyramid of adversarial networks, in: Advances in neural information processing systems, 2015, pp. 1486–1494.
- [111] S. R. Miller, C. S. Tucker, Human validation of computer vs human generated design sketches.
- [112] M. S. Silberman, K. Milland, R. LaPlante, Stop citing ross et al. 2010,’who are the crowdworkers?’, Retrieved August 28 (2015) 2016.
- [113] Z. Wang, A. C. Bovik, Mean squared error: Love it or leave it? a new look at signal fidelity measures, IEEE signal processing magazine 26 (1) (2009) 98–117.
- [114] V. Dumoulin, I. Belghazi, B. Poole, A. Lamb, M. Arjovsky, O. Mastropietro, A. Courville, Adversarially learned inference, arXiv preprint arXiv:1606.00704.
- [115] L. Metz, B. Poole, D. Pfau, J. Sohl-Dickstein, Unrolled generative adversarial networks, arXiv preprint arXiv:1611.02163.
- [116] A. Oliva, Gist of the scene, in: Neurobiology of attention, Elsevier, 2005, pp. 251–256.
- [117] T. Serre, A. Oliva, T. Poggio, A feedforward architecture accounts for rapid categorization, Proceedings of the national academy of sciences 104 (15) (2007) 6424–6429.
- [118] J. Snell, K. Ridgeway, R. Liao, B. D. Roads, M. C. Mozer, R. S. Zemel, Learning to generate images with perceptual similarity metrics, arXiv preprint arXiv:1511.06409.
- [119] P. Upchurch, J. Gardner, K. Bala, R. Pless, N. Snavely, K. Q. Weinberger, Deep feature interpolation for image content changes, arXiv preprint arXiv:1611.05507.
- [120] C. Donahue, A. Balsubramani, J. McAuley, Z. C. Lipton, Semantically decomposing the latent spaces of generative adversarial networks, arXiv preprint arXiv:1705.07904.
- [121] Y. Liu, Z. Qin, Z. Luo, H. Wang, Auto-painter: Cartoon image generation from sketch by using conditional generative adversarial networks, arXiv preprint arXiv:1705.01908.
- [122] Y. Lu, S. Wu, Y.-W. Tai, C.-K. Tang, Sketch-to-image generation using deep contextual completion, arXiv preprint arXiv:1711.08972.
- [123] Z. C. Lipton, S. Tripathi, Precise recovery of latent vectors from generative adversarial networks, arXiv preprint arXiv:1702.04782.

- [124] D. Berthelot, T. Schumm, L. Metz, Began: Boundary equilibrium generative adversarial networks, arXiv preprint arXiv:1703.10717.
- [125] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, J. Clune, Synthesizing the preferred inputs for neurons in neural networks via deep generator networks, in: Advances in Neural Information Processing Systems, 2016, pp. 3387–3395.
- [126] T. White, Sampling generative networks: Notes on a few effective techniques, arXiv preprint arXiv:1609.04468.
- [127] R. Vedantam, I. Fischer, J. Huang, K. Murphy, Generative models of visually grounded imagination, arXiv preprint arXiv:1705.10762.
- [128] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Object detectors emerge in deep scene cnns, arXiv preprint arXiv:1412.6856.
- [129] J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net, arXiv preprint arXiv:1412.6806.
- [130] L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, Journal of machine learning research 9 (Nov) (2008) 2579–2605.
- [131] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, E. Shechtman, Toward multimodal image-to-image translation, in: Advances in Neural Information Processing Systems, 2017, pp. 465–476.
- [132] D. J. Im, A. H. Ma, G. W. Taylor, K. Branson, Quantitatively evaluating gans with divergences proposed for training, International Conference on Learning RepresentationsAccepted as poster.
URL <https://openreview.net/forum?id=SJQHjzZ0->
- [133] H. E. Gerhard, F. A. Wichmann, M. Bethge, How sensitive is the human visual system to the local statistics of natural images?, PLoS computational biology 9 (1) (2013) e1002873.
- [134] J. Driver, G. C. Baylis, R. D. Rafal, Preserved figure-ground segregation and symmetry perception in visual neglect, Nature 360 (6399) (1992) 73.
- [135] C. Funk, Y. Liu, Beyond planar symmetry: Modeling human perception of reflection and rotation symmetries in the wild, arXiv preprint arXiv:1704.03568.
- [136] R. Rosenholtz, Y. Li, L. Nakano, Measuring visual clutter, Journal of vision 7 (2) (2007) 17–17.