

Reliable Fidelity and Diversity Metrics for Generative Models

Muhammad Ferjad Naeem^{*1,2} Seong Joon Oh^{*1} Youngjung Uh¹ Yunjey Choi¹ Jaejun Yoo^{1,3}

Abstract

Devising indicative evaluation metrics for the image generation task remains an open problem. The most widely used metric for measuring the similarity between real and generated images has been the Fréchet Inception Distance (FID) score. Because it does not differentiate the *fidelity* and *diversity* aspects of the generated images, recent papers have introduced variants of precision and recall metrics to diagnose those properties separately. In this paper, we show that even the latest version of the precision and recall metrics are not reliable yet. For example, they fail to detect the match between two identical distributions, they are not robust against outliers, and the evaluation hyperparameters are selected arbitrarily. We propose **density and coverage** metrics that solve the above issues. We analytically and experimentally show that density and coverage provide more interpretable and reliable signals for practitioners than the existing metrics. Code: github.com/clovaai/generative-evaluation-prdc.

1. Introduction

Assessing a generative model is difficult. Unlike the evaluation of discriminative models $P(T|X)$ that is often easily done by measuring the prediction performances on a few labelled samples (X_i, T_i) , generative models $P(X)$ are assessed by measuring the discrepancy between the real $\{X_i\}$ and generated (fake) $\{Y_j\}$ sets of high-dimensional data points. Adding to the complexity, there are more than one way of measuring distances between two distributions each with its own pros and cons. In fact, even human judgement based measures like Mean Opinion Scores (MOS) are not ideal, as practitioners have diverse opinions on what the “ideal” generative model is (Borji, 2019).

^{*}Equal contribution ¹Clova AI Research, Naver Corp.

²Technische Universität München, Germany ³École polytechnique fédérale de Lausanne (EPFL), Switzerland. Correspondence to: Jaejun Yoo <jaejun.yoo88@gmail.com>.

Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by the author(s).

Nonetheless, there must be some measurement of the quality of generative models for the progress of science. Several quantitative metrics have been proposed, albeit with their own set of trade-offs. For example, Fréchet Inception Distance (FID) score (Heusel et al., 2017), the most popular metric in image generation tasks, has empirically exhibited good agreements with human perceptual scores. However, FID summarises the comparison of two distributions into a single number, failing to separate two important aspects of the quality of generative models: **fidelity and diversity** (Sajjadi et al., 2018). Fidelity refers to the degree to which the generated samples resemble the real ones. Diversity, on the other hand, measures whether the generated samples cover the full variability of the real samples.

Recent papers (Sajjadi et al., 2018; Simon et al., 2019; Kynkänniemi et al., 2019) have introduced precision and recall metrics as measures of fidelity and diversity, respectively. Though precision and recall metrics have introduced the important perspectives in generative model evaluation, we show that they are not ready yet for practical use. We argue that necessary conditions for useful evaluation metrics are: (1) ability to detect identical real and fake distributions, (2) robustness to outlier samples, (3) responsiveness to mode dropping, and (4) the ease of hyperparameter selection in the evaluation algorithms. Unfortunately, even the most recent version of the precision and recall metrics (Kynkänniemi et al., 2019) fail to meet the requirements.

To address the practical concerns, we propose the **density and coverage** metrics. By introducing a simple yet carefully designed manifold estimation procedure, we not only make the fidelity-diversity metrics empirically reliable but also theoretically analysable. We test our metric on generative adversarial networks, one of the most successful generative models in recent years.

We then study the embedding algorithms for evaluating image generation algorithms. Embedding is an inevitable ingredient due to the high-dimensionality of images and the lack of semantics in the RGB space. Despite the importance, the embedding pipeline has been relatively less studied in the existing literature; evaluations of generated images mostly rely on the features from an ImageNet pre-trained model (Salimans et al., 2016; Heusel et al., 2017; Sajjadi et al., 2018; Simon et al., 2019; Kynkänniemi et al.,

2019; Deng et al., 2009). This sometimes limits the fair evaluation and provides a false sense of improvement, since the pre-trained models inevitably include the dataset bias (Torralba & Efros, 2011; Geirhos et al., 2019). We show that such pre-trained embeddings often exhibit unexpected behaviours as the target distribution moves away from the natural image domain.

To exclude the dataset bias, we consider using randomly initialised CNN feature extractors (Ulyanov et al., 2018). We compare the evaluation metrics on MNIST and sound generation tasks using the random embeddings. We observe that random embeddings provide more macroscopic views on the distributional discrepancies. In particular, random embeddings provide more sensible evaluation results when the target data distribution is significantly different from ImageNet statistics (*e.g.* MNIST and spectrograms).

2. Backgrounds

Given a real distribution $P(X)$ and a generative model $Q(Y)$, we assume that we can sample $\{X_i\}$ and $\{Y_j\}$, respectively. We need an algorithm to assess how likely the sets of samples are arising from the same distribution. When the involved distribution families are tractable and the full density functions can be easily estimated, statistical testing methods or distributional distance measures (*e.g.* Kullback-Leibler Divergence or Expected Likelihood) are viable. However, when the data $P(X)$ are complex and high-dimensional (*e.g.* natural images), it becomes difficult to apply such measures naively (Theis et al., 2016). Because of the difficulty, the evaluation of samples from generative models is still an actively researched topic. In this section, we provide an overview of existing approaches. We describe the most widely-used evaluation pipeline for image generative models (§2.1), and then introduce the prior works on fidelity and diversity measures (§2.2). For a more extensive survey, see (Borji, 2019).

2.1. Evaluation pipeline

It is difficult to conduct statistical analyses over complex and high-dimensional data X in their raw form (*e.g.* images). Thus, evaluation metrics for image generative models largely follow the following stages: (1) embed real and fake data ($\{X_i\}$ and $\{Y_j\}$) into a Euclidean space \mathbb{R}^D through a non-linear mapping f like CNN feature extractors, (2) construct real and fake distributions over \mathbb{R}^D with the embedded samples $\{f(X_i)\}$ and $\{f(Y_j)\}$, and (3) quantify the discrepancy between the two distributions. We describe each stage in the following paragraphs.

Embeddings. It is often difficult to define a sensible metric over the input space. For example, ℓ_2 distance over the image pixels $\|X_i - Y_j\|_2$ is misleading because two

perceptually identical images may have great ℓ_2 distances (one-pixel translation) (Theis et al., 2016). To overcome this difficulty, researchers have introduced ImageNet pre-trained CNN feature extractors as the embedding function f in many generative model evaluation metrics (Salimans et al., 2016; Heusel et al., 2017; Sajjadi et al., 2018; Kynkänniemi et al., 2019) based on the reasoning that the ℓ_2 distance in the feature space $\|f(X_i) - f(Y_j)\|_2$ provide sensible proxies for the human perceptual metric (Zhang et al., 2018). Since we always use embedded samples for computing metrics, we write X_i and Y_j for $f(X_i)$ and $f(Y_j)$, respectively.

In this work, we also adopt ImageNet pre-trained CNNs for the embedding, but we also criticise their use when the data distribution is distinct from the ImageNet distribution (§4.2). We suggest randomly-initialised CNN feature extractors as an alternative in such cases (Ulyanov et al., 2018; Zhang et al., 2018). We show that for MNIST digits or sound spectrograms with large domain gaps from ImageNet, random embeddings provide more sensible evaluation measures.

Building and comparing distributions. Given embedded samples $\{X_i\}$ and $\{Y_j\}$, many metrics conduct some form of (non-)parametric statistical estimation. *Parzen window estimates* (Bengio et al., 2013) approximate the likelihoods of the fake samples $\{Y_j\}$ by estimating the density $P(X)$ with Gaussian kernels around the real samples $\{X_i\}$. On the parametric side, *Inception scores (IS)* (Salimans et al., 2016) estimate the multinomial distribution $P(T|Y_j)$ over the 1000 ImageNet classes for each sample image Y_j and compares it against the estimated marginalised distribution $P(T)$ with the KL divergence. *Fréchet Inception Distance (FID)* (Heusel et al., 2017) estimates the mean μ and covariance Σ for X and Y assuming that they are multivariate Gaussians. The distance between the two Gaussians is computed by the Fréchet distance (Dowson & Landau, 1982), also known as the Wasserstein-2 distance (Vaserstein, 1969). FID has been reported to generally match with human judgements (Heusel et al., 2017); it has been the most popular metric for image generative models in the last couple of years (Borji, 2019).

2.2. Fidelity and diversity

While single-value metrics like IS and FID have led interesting advances in the field by ranking generative models, they are not ideal for diagnostic purposes. One of the most important aspects of the generative models is the trade-off between *fidelity* (how realistic each input is) and *diversity* (how well fake samples capture the variations in real samples). We introduce variants of the two-value metrics (precision and recall) that capture the two characteristics separately.

Precision and recall. (Sajjadi et al., 2018) have reported the pathological case where two generative models have similar FID scores, while their qualitative fidelity and diver-

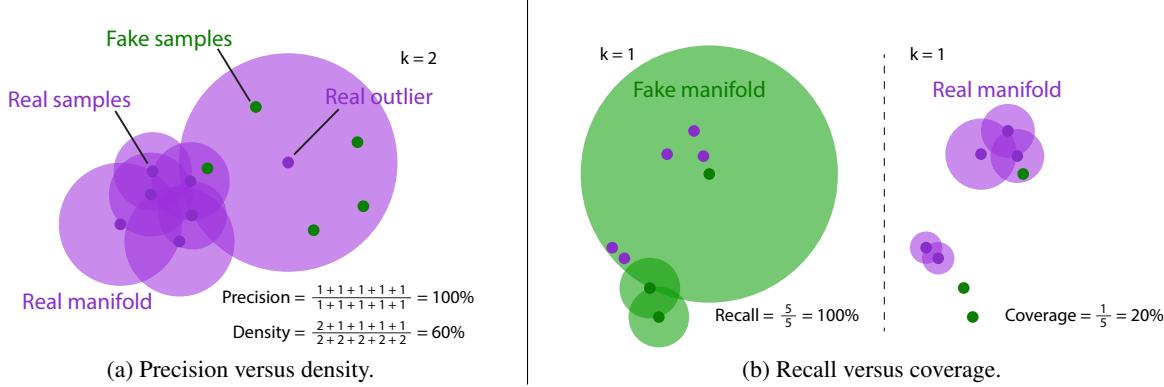


Figure 1. Overview of metrics. Two example scenarios for illustrating the advantage of density over precision and coverage over recall. Note that for recall versus coverage figure, the real and fake samples are identical across left and right.

sity results are different. For a better diagnosis of generative models, (Sajjadi et al., 2018) have thus proposed the *precision and recall* metrics based on the estimated supports of the real and fake distributions. Precision is defined as the portion of $Q(Y)$ that can be generated by $P(X)$; recall is symmetrically defined as the portion of $P(X)$ that can be generated by $Q(Y)$. While conceptually useful, they have multiple practical drawbacks. It assumes that the embedding space is uniformly dense, relies on the initialisation-sensitive k-means algorithm for support estimation, and produces an infinite number of values as the metric.

Improved precision and recall. (Kynkänniemi et al., 2019) have proposed the *improved precision and recall* (P&R) that address the above drawbacks. The probability density functions are estimated via k-nearest neighbour distances, overcoming the uniform-density assumption and the reliance on the k-means. Our proposed metrics are based on P&R; we explain the full details of P&R here.

P&R first constructs the “manifold” for $P(X)$ and $Q(Y)$ separately, the object is nearly identical to the probabilistic density function except that it does not sum to 1. Precision then measures the expected likelihood of fake samples against the real manifold and recall measures the expected likelihood of real samples against the fake manifold:

$$\text{precision} := \frac{1}{M} \sum_{j=1}^M \mathbb{1}_{Y_j \in \text{manifold}(X_1, \dots, X_N)} \quad (1)$$

$$\text{recall} := \frac{1}{N} \sum_{i=1}^N \mathbb{1}_{X_i \in \text{manifold}(Y_1, \dots, Y_M)} \quad (2)$$

where N and M are the number of real and fake samples. $\mathbb{1}(\cdot)$ is the indicator function. Manifolds are defined

$$\text{manifold}(X_1, \dots, X_N) := \bigcup_{i=1}^N B(X_i, \text{NND}_k(X_i)) \quad (3)$$

where $B(x, r)$ is the sphere in \mathbb{R}^D around x with radius r . $\text{NND}_k(X_i)$ denotes the distance from X_i to the k^{th} nearest neighbour among $\{X_i\}$ excluding itself. Example computation of P&R is shown in Figure 1.

There are similarities between the precision above and the Parzen window estimate (Bengio et al., 2013). If the manifolds are formed by *superposition* instead of *union* in Equation 3 and the spheres $B(\cdot, \cdot)$ are replaced with Gaussians of fixed variances, then the manifold estimation coincides with the kernel density estimation. In this case, Equation 1 computes the expected likelihood of fake samples.

3. Density and Coverage

We propose novel performance measures **density and coverage** (**D&C**) as practically usable measures that successfully remedy the problems with precision and recall.

3.1. Problems with improved precision and recall

Practicality of the improved precision and recall (P&R) is still compromised due to their vulnerability to outliers and computational inefficiency. Building the nearest neighbour manifolds (Equation 3) must be performed carefully because the spheres around each sample are not normalised according to the their radii or the relative density of samples in the neighbourhood. Consequently, the nearest neighbour manifolds generally overestimate the true manifold around the outliers, leading to undesired effects in practice. We explain the drawbacks at the conceptual level here; they will be quantified in §4.

Precision. We first show a pathological case for precision in Figure 1a. Because of the real outlier sample, the manifold is overestimated. Generating many fake samples around the real outlier is enough to increase the precision measure.

Recall. The nearest neighbour manifold is built upon the

fake samples in Equation 2. Since models tend to generate many unrealistic yet diverse samples, the fake manifold is often an overestimation of the actual fake distribution. The pathological example is shown in Figure 1b. While the fake samples are far from the modes in real samples, the recall measure gains points for real samples contained in the overestimated fake manifold. Another problem with relying on fake manifolds for the recall computation is that the manifold must be computed per model. For example, to generate the recall-vs-iteration curve for training diagnosis, the k-nearest neighbours for all fake samples must be computed ($O(kM \log M)$) for every data point.

3.2. Density and coverage

We remedy the issues with P&R above and propose new metrics: density and coverage (D&C).

Density. *Density* improves upon the precision metric by fixing the overestimation of the manifold around real outliers. Precision counts the binary decision of whether the fake data Y_j contained in *any* neighbourhood sphere $\{B(X_i, \text{NND}_k(X_i))\}_i$ of real samples (Equation 1). Density, instead, counts *how many* real-sample neighbourhood spheres contain Y_j (Equation 4). The manifold is now formed by the superposition of the neighbourhood spheres $\{B(X_i, \text{NND}_k(X_i))\}_i$, and a form of expected likelihood of fake samples is measured. In this sense, the density measure is at the midway between the precision metric and the Parzen window estimate. Density is defined as

$$\text{density} := \frac{1}{kM} \sum_{j=1}^M \sum_{i=1}^N 1_{Y_j \in B(X_i, \text{NND}_k(X_i))} \quad (4)$$

where k is for the k-nearest neighbourhoods. Through this modification, density rewards samples in regions where real samples are densely packed, relaxing the vulnerability to outliers. For example, the problem of overestimating precision (100%) in Figure 1a is resolved using the density measure (60%). Note that unlike precision, density is not upper bounded by 1; it may be greater than 1 depending on the density of reals around the fakes.

Coverage. Diversity, intuitively, shall be measured by the ratio of real samples that are covered by the fake samples. *Coverage* improves upon the recall metric to better quantify this by building the nearest neighbour manifolds around the real samples, instead of the fake samples, as they have less outliers. Moreover, the manifold can only be computed per dataset, instead of per model, reducing the heavy nearest neighbour computations in recall. Coverage is defined as

$$\text{coverage} := \frac{1}{N} \sum_{i=1}^N 1_{\exists j \text{ s.t. } Y_j \in B(X_i, \text{NND}_k(X_i))}. \quad (5)$$

It measures the fraction of real samples whose neighbour-

hoods contain at least one fake sample. Coverage is bounded between 0 and 1.

3.3. Analytic behaviour of density & coverage

The simplest sanity check for an evaluation metric is whether the metric attains the best value when the intended criteria are met. For generative models, we examine if D&C attain 100% performances when the real and fake distribution are identical ($P \stackrel{d}{=} Q$). We show that, unlike P&R, D&C yield an analytic expression for the expected values $\mathbb{E}[\text{density}]$ and $\mathbb{E}[\text{coverage}]$ for identical real and fake, and the values approach 100% as the sample sizes (N, M) and number of neighbourhoods k increase. This analysis further leads to a systematic algorithm for selecting the hyperparameters (k, N, M).

Lack of analytic results for P&R. From Equation 1, the expected precision for identical real and fake is

$$\mathbb{E}[\text{precision}] = \mathbb{P}[X_0 \in \text{manifold}(X_1, \dots, X_N)] \quad (6)$$

$$= \mathbb{P}\left[X_0 \in \bigcup_{i=1}^N B(X_i, \text{NND}_k(X_i))\right] \quad (7)$$

$$= \mathbb{P}\left[\bigcup_{i=1}^N A_i^k\right] \quad (8)$$

where A_i^k is the event $\{\|X_0 - X_i\| < \text{NND}_k(X_i)\}$. Since $(A_i^k)_i$ are not independent with complex dependence structures, a simple expression for Equation 8 does not exist. Same observation holds for $\mathbb{E}[\text{recall}]$.

Analytic derivations for D&C. We derive the expected values of D&C under the identical real and fake.

Lemma 1. $\mathbb{E}[\text{density}] = 1$.

Proof. The expected density boils down to

$$\mathbb{E}[\text{density}] = \frac{1}{k} \sum_{i=1}^N \mathbb{P}(B_N^k) = \frac{N}{k} \mathbb{P}(B_N^k)$$

where B_N^k is the event where $\|Y - X_1\|$ is at most k^{th} smallest among the random variables $\mathbf{S} := \{\|X_1 - X_2\|, \dots, \|X_1 - X_N\|\}$. Since the random variables $\mathbf{S} \cup \{\|Y - X_1\|\}$ are identical and independently distributed, any particular ranking of them are equally likely with probability $1/N!$. Since the number of rankings of N values with one of them at a particular rank is $(N-1)!$, we have

$$\mathbb{P}(B_N^k) = k \cdot (N-1)! \cdot \frac{1}{N!} = \frac{k}{N}. \quad \blacksquare$$

Lemma 2.

$$\mathbb{E}[\text{coverage}] = 1 - \frac{(N-1) \cdots (N-k)}{(M+N-1) \cdots (M+N-k)}. \quad (9)$$

Moreover, as $M = N \rightarrow \infty$, $\mathbb{E}[\text{coverage}] \rightarrow 1 - \frac{1}{2^k}$.

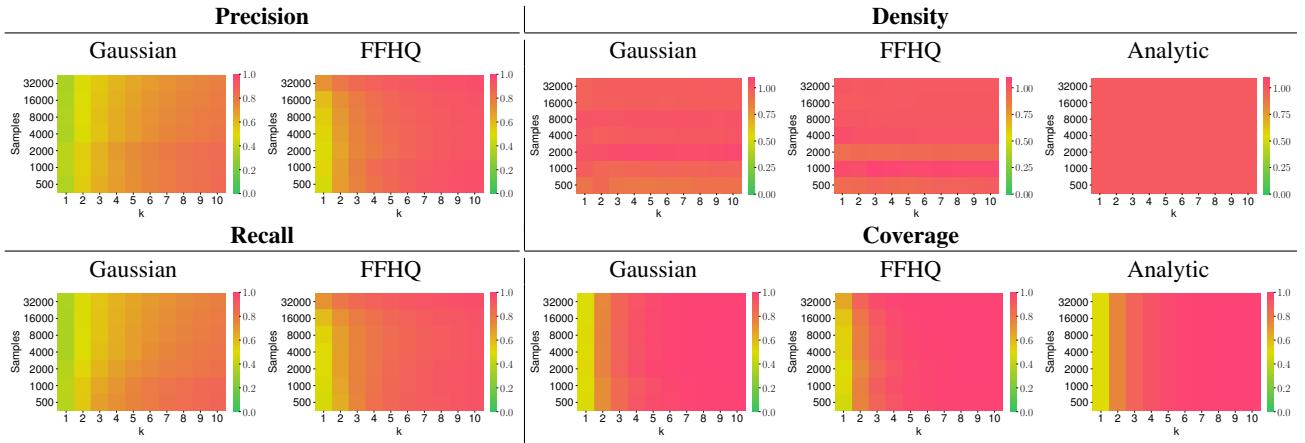


Figure 2. Metrics under identical real and fake distributions. Results with Gaussian and FFHQ data are shown. For density and coverage, the expected values derived in §3.3 are plotted. In each case, the values are shown with respect to the varying number of samples $N = M$ (assume the same number of real and fake samples) and the nearest neighbours k .

Proof. The expected coverage is estimated as

$$\begin{aligned}\mathbb{E}[\text{coverage}] &= \frac{1}{N} \sum_{i=1}^N \mathbb{P}(\exists j \text{ s.t. } Y_j \in B(X_i, \text{NND}_k(X_i))) \\ &\quad (*) \\ &= 1 - \mathbb{P}(\forall j, Y_j \notin B(X_1, \text{NND}_k(X_1))) \\ &\quad (**)\end{aligned}$$

using the fact that the events $(*)$ is symmetrical with respect to i . The event $(**)$ can be re-written as

$$\min_j \|Y_j - X_1\| \geq \|X_1 - X_\beta\| \text{ for at most } k \beta's.$$

We write $Z_\beta := \|X_1 - X_\beta\|$ for $\beta \in \{2, \dots, N\}$ and $\tilde{Z}_j := \|X_1 - Y_j\|$ for $j \in \{1, \dots, M\}$. Then, the set $\{Z_\beta\}_{\beta=2}^N \cup \{\tilde{Z}_j\}_{j=1}^M$ of $N + M - 1$ random variables is independent and identically distributed. The probability of $(**)$ can be equivalently described as:

Assume there are $M + N - 1$ non-negative real numbers Z distributed according to $\overset{\text{iid}}{\mathbb{P}}$. Colour M of them red uniformly at random and colour the rest $N - 1$ blue. What is the chance that the $k - 1$ smallest among Z are all coloured blue?

Since any assignment of red and blue colours is equally likely, we compute the probability by counting the ratio of possible colour assignments where k smallest elements are coloured blue. The formula is written as

$$\frac{\binom{M+N-k-1}{M}}{\binom{M+N-1}{M}} = \frac{(N-1) \cdots (N-k)}{(M+N-1) \cdots (M+N-k)}.$$

3.4. Hyperparameter selection

Given the analytic expression for the expected D&C for identical real and fake distributions, we can systematically choose k , M , and N . We set the aim of hyperparameter selection as: $\mathbb{E}[\text{coverage}] > 1 - \epsilon$ for identical real and fake.

Since $\mathbb{E}[\text{coverage}]$ does not depend upon the exact distribution type or dimensionality of data (Equation 9), the hyperparameters chosen as above will be effective across various data types. We verify the consistency of D&C across data types in Figure 2. It contains plots of the P&R and D&C values for identical real and fake distributions from (1) 64-dimensional multivariate standard Gaussians, (2) 4096-dimensional ImageNet pre-trained VGG embeddings of the FFHQ face images, and (3) analytic estimations in §3.3. While P&R exhibits significantly different estimated values across the Gaussians and the FFHQ embeddings, D&C metrics agree on all three types of estimates, confirming the independence on data type. This conceptual advantage leads to a confident choice of evaluation hyperparameters that are effective across data types.

In practice, we choose the hyperparameters to achieve $\mathbb{E}[\text{coverage}] > 0.95$. For the sake of symmetry, we first set $M = N$. We then set $M = N = 10\,000$ to ensure a good approximation $\mathbb{E}[\text{coverage}] \approx 1 - \frac{1}{2^k}$, while keeping the computational cost tractable. $k = 5$ is then sufficient to ensure $\mathbb{E}[\text{coverage}] \approx 0.969 > 0.95$. The exact principle behind the choices of those values for P&R ($k = 3$, $M = N = 50\,000$) is unknown.

4. Experiments

We empirically assess the proposed density and coverage (D&C) metrics and compare against the improved precision and recall (P&R) (Kynkänniemi et al., 2019). Evaluating

Note that the expected values do not depend upon the distribution type or the dimensionality D of the data.

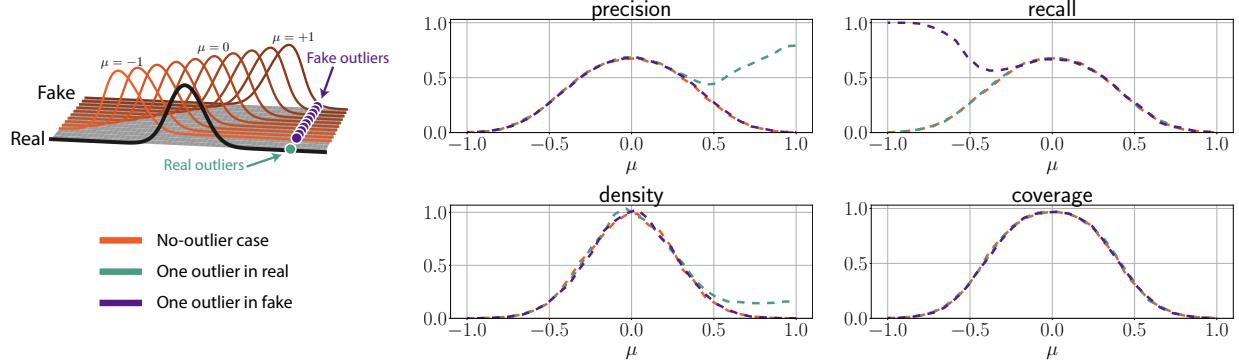


Figure 3. Robustness to outliers for toy data. Behaviour of the four metrics when the real distribution is fixed $X \sim N(0, I)$ and fake distribution $Y \sim N(\mu, I)$ shifts with $\mu \in [-1, 1]$. We further consider two outlier scenarios where a sample at $x = +3$ is added either to the set of real or fake samples.

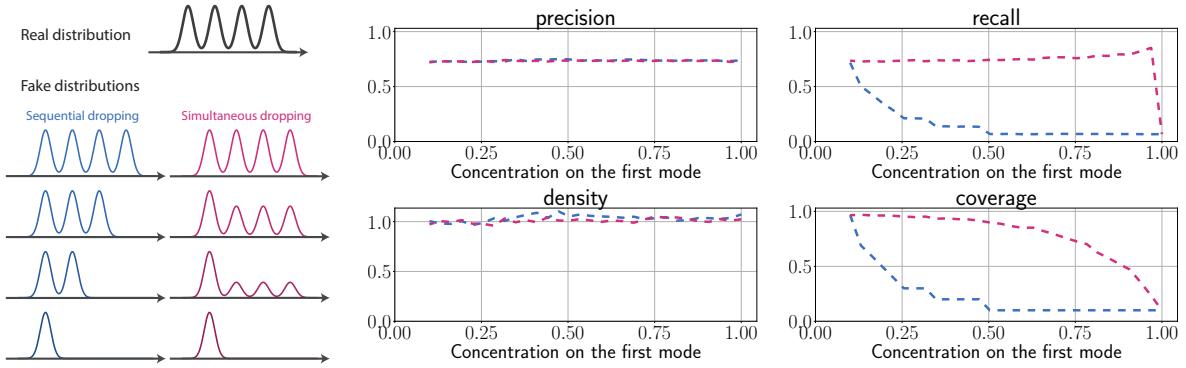


Figure 4. Sensitivity to mode dropping for toy data. Assuming the mixture-of-Gaussian real distribution X , we simulate two paths in which the perfect fake distribution Y drops to a single mode. Under the *sequential dropping*, Y drops a mode one at a time; under the *simultaneous dropping*, Y gradually decreases samples in all but one mode. Corresponding behaviours of the four metrics are shown.

evaluation metrics is difficult as the ground truth metric values are often unavailable (which is why the evaluation metrics are proposed in the first place). We carefully select sanity checks with toy and real-world data where the desired behaviours of evaluation metrics are clearly defined. At the end of the section, we study the embedding pipeline (§2.1) and advocate the use of randomly initialised embedding networks under certain scenarios.

4.1. Empirical results on density and coverage

We build several toy and real-world data scenarios to examine the behaviour of the four evaluation metrics: P&R and D&C. We first show results on toy data where the desired behaviours of the fidelity and diversity metrics are well-defined (§4.1.1). We then move on to diverse real-world data cases to show that the observations extend to complex data distributions (§4.1.2).

4.1.1. SANITY CHECKS WITH TOY DATA

We assume Gaussian or mixture-of-Gaussian distributions for the real X and fake Y in \mathbb{R}^{64} ($D = 64$). We simulate

largely two scenarios: (1) Y moves away from X (Figure 3) and (2) Y gradually fails to cover the modes in X (Figure 4). We discuss each result in detail below.

Translating Y away from X . We set $X \sim N(0, I)$ and $Y \sim N(\mu \mathbf{1}, I)$ in \mathbb{R}^{64} where $\mathbf{1}$ is the vector of ones and I is the identity matrix. We study how the metrics change as μ varies in $[-1, 1]$. In Figure 3, without any outlier, this setting leads to the decreasing values of P&R and D&C as μ moves away from 0, the desired behaviour for all metrics. However, P&R show a pathological behaviour when the distributions match ($\mu = 0$): their values are far below 1 (0.68 precision and 0.67 recall). D&C, on the other hand, achieve values close to 1 (1.06 density and 0.97 coverage). D&C detect the distributional match better than P&R.

Translating with real or fake outliers. We repeat the previous experiment with exactly one outlier at $\mathbf{1} \in \mathbb{R}^{64}$ in either real or fake samples (Figure 3). Robust metrics must not be affected by one outlier sample. However, P&R are vulnerable to this one sample; precision increases as μ grows above 0.5 and recall increases as μ decreases below -0.4 . The behaviour is attributed to the overestimation

of the manifold on the region enclosed by the inliers and the outlier (§3.1). This susceptibility is a serious issue in practice because outliers are common in realistic data and generated outputs. On the other hand, the D&C measures for the outlier cases largely coincide with the no-outlier case. D&C are sturdy measures.

Mode dropping. We assume that the real distribution X is a mixture of Gaussians in \mathbb{R}^{64} with ten modes. We simulate the fake distribution Y initially as identical to X , and gradually drop all but one mode. See Figure 4 for an illustration. We consider two ways the modes are dropped. (1) Each mode is dropped sequentially and (2) weights on all but one mode are decreased simultaneously. Under the sequential mode dropping, both recall and coverage gradually drop, capturing the decrease in diversity. However, under the simultaneous dropping, recall cannot capture the decrease in diversity until the concentration on the first mode reaches 90% and shows a sudden drop when the fake distribution becomes unimodal. Coverage, on the other hand, decreases gradually even under the simultaneous dropping. It reliably captures the decrease in diversity in this case.

4.1.2. SANITY CHECKS WITH REAL-WORLD DATA

Having verified the metrics on toy Gaussians, we assess the metrics on real-world images. As in the toy experiments, we focus on the behaviour of metrics under corner cases including outliers and mode dropping. We further examine the behaviour with respect to the latent truncation threshold ψ (Karras et al., 2019). As the embedding network, we use the f_{c2} layer features of the ImageNet pre-trained VGG16 (Kynkänniemi et al., 2019).

Outliers. Before studying the impact of outliers on the evaluation metrics, we introduce our criterion for outlier detection. Motivated by (Kynkänniemi et al., 2019), we use the distance to the k^{th} nearest neighbour among the fake samples. According to this criterion, we split the fake samples into 10 : 1 for inliers and outliers. We experiment with fake images from StyleGAN on CelebA (Liu et al., 2015) and LSUN-bedroom (Yu et al., 2015). Example images of inliers and outliers are shown in Figure 5. We observe that the outliers have a more distortions and atypical semantics.

We examine the behaviour of recall and coverage as the outliers are gradually added to the pool of fake samples. In Figure 5, we plot recall and coverage relative to their values when there are only inliers. As outliers are added, recall increases more than 11% and 15% on CelebA and LSUN bedroom, respectively, demonstrating its vulnerability to outliers. Coverage, on the other hand, is stable: less than 2% increase with extra outliers.

Mode dropping. As in the toy experiments, we study mode dropping on the MNIST digit images. We treat the ten

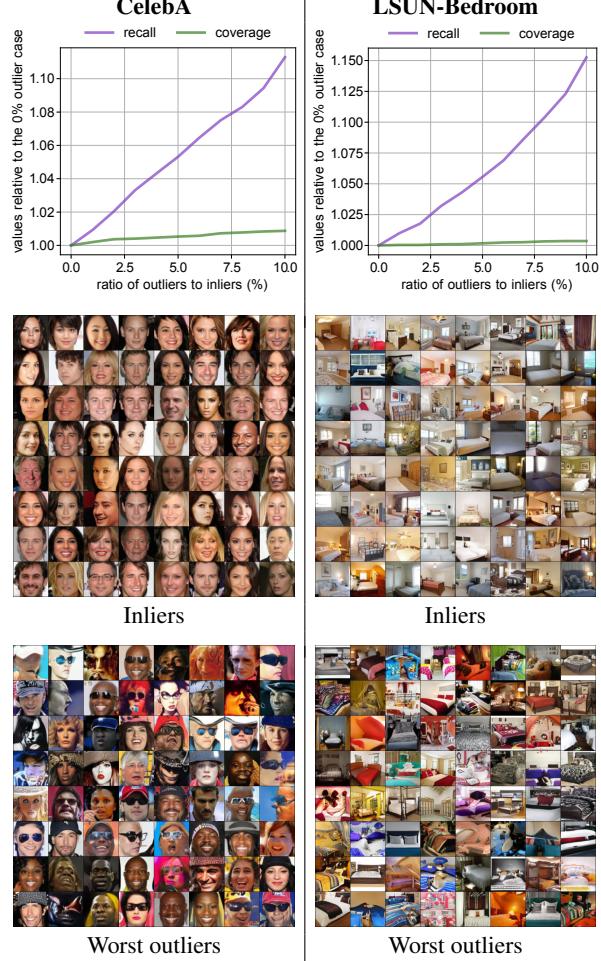


Figure 5. Recall versus coverage against the amount of outliers. Example inliers and the worst outliers are shown below.

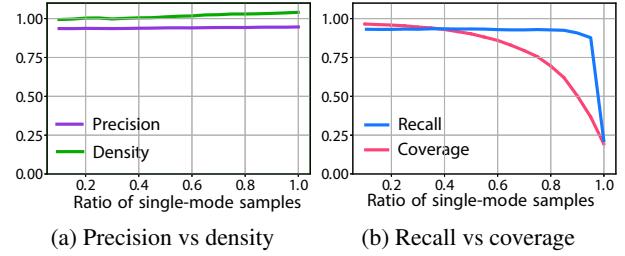


Figure 6. Metrics under mode dropping on MNIST. Behaviour of the metrics under varying degrees of the mode dropping towards a single class on MNIST with R^{64} . The real distribution has the uniform class distribution while the fake distribution has an increasing ratio of the class “0”.

classes as modes and simulate the scenario where a generative model gradually favours a particular mode (class “0”) over the others. We use the real data (MNIST images) with decreasing number of classes covered as our fake samples (sequential dropping). The results are shown in Figure 6.

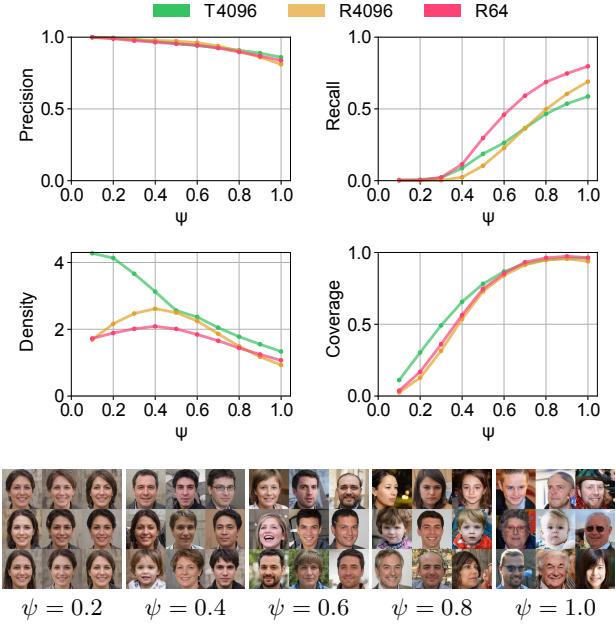


Figure 7. Behaviour of metrics with ψ . The latent truncation threshold ψ is applied over over StyleGAN generated images on FFHQ. Qualitative examples are given.

While recall is unable to detect the decrease in overall diversity until the ratio of class “0” is over 90%, coverage exhibits a gradual drop. Coverage is superior to recall at detecting mode dropping.

Resolving fidelity and diversity. The main motivation behind two-value metrics like P&R is the diagnosis involving the fidelity and diversity of generated images. We validate whether D&C successfully achieve this. We perform the truncation analysis on StyleGAN (Karras et al., 2019) on FFHQ. The truncation technique is used in generative models to artificially manipulate the learned distributions by thresholding the latent noise Z with $\psi \geq 0$ (Brock et al., 2018). In general, for greater ψ , data fidelity decreases and diversity increases, and vice versa. This stipulates the desired behaviours of the P&R and D&C metrics.

The results are shown in Figure 7. With increasing ψ , precision and density decrease, while recall and coverage increase. Note that density varies more than precision as ψ increases, leading to finer-grained diagnosis for the fidelity.

Relation to negative log likelihood metrics. For variational auto-encoders (VAE, (Kingma & Welling, 2013)) and their variants, negative log likelihood (NLL) has been a popular metric for quantifying the ability of the trained model to represent and replicate the data distribution. While NLL provides a single view of the generative model, the D&C metrics provide further diagnostic information. For example, we have trained a convolutional VAE with MNIST. The mean NLL is computed on the reconstructed test data; the D&C

Real MNIST			Real spectrogram		
1	7	9	9	5	6
3	2	3	7	4	5
6	3	0	3	1	6
0	7	5	4	1	2
			1	0	6
			3	2	4
			5	6	2
			1	0	2
			2	6	5
			3	3	9
			8	8	0
			7	3	4
			6	6	4
			0	8	0
DCGAN			WaveGAN		
1	3	9	6	7	2
5	6	2	4	9	3
1	0	2	6	5	3
7	3	4	6	4	1
5	6	4	9	0	8
1	0	2	6	5	3
2	6	5	3	9	8
3	3	9	8	0	0
4	1	1	1	1	1
5	6	2	4	9	3
6	6	4	9	0	8
7	3	4	6	4	1
8	8	0	0	0	0
9	9	5	0	0	0

Table 1. Failure of trained embeddings. Evaluation results of generated samples beyond natural images using T4096, R4096, and R64. Corresponding real and fake qualitative samples shown.

are computed over randomly-initialised VGG features of dimension 64 (§4.2). To simulate the fidelity-diversity trade-off for VAEs, we consider truncating the latent variable z at magnitude 1. The NLL for vanilla $-\log p(x_{\text{test}}|z)$ versus truncated latent $-\log p(x_{\text{test}}|z, \|z\| \leq 1)$ is 41.4 and 40.2, not providing meaningful granularity on fidelity-diversity trade-off. For the same setup, D&C provide further details: density has not changed much (0.208→0.205), while coverage drops significantly (0.411→0.291) as the result of truncation. Latent variable truncation for our VAE model leads to a drop in diversity, while not improving the fidelity.

4.2. Random Embeddings

Image embedding is an important component in generative model evaluation (§2.1). Yet, this ingredient is relatively less studied in existing literature. In this section, we explore the limitations of the widely-used ImageNet embeddings when the target data is distinct from the ImageNet samples. In such cases, we propose random embeddings as alternatives. In the following experiments, we write T4096 for the f_{c2} features of VGG16 pre-trained on ImageNet, R4096 for the f_{c2} features of a randomly initialised VGG16, and R64 for the $f_{c2'}$ features of a randomly initialised VGG16 where the prime indicates the replacement of the 4096 dimensions with 64 dimensions. We experiment on MNIST and a sound dataset, Speech Commands Zero Through Nine (SC09).

4.2.1. METRICS WITH R64 ON NATURAL IMAGES

We study the behaviour of metrics for different truncation thresholds (ψ) on StyleGAN generated FFHQ images (Fig-

ure 7). We confirm that the metrics on R64 closely follow the trend with T4096. P&R and D&C over R64 can be used in ImageNet-like images to capture the fidelity and diversity aspects of model outputs.

4.2.2. METRICS WITH R64 BEYOND NATURAL IMAGES

We consider the scenario where the target distribution is significantly different from the ImageNet statistics. We use DCGAN generated images (Radford et al., 2016) on MNIST (LeCun et al., 1998) and WaveGAN generated spectrograms (Donahue et al., 2019) on Speech Commands Zero Through Nine (SC09) dataset. The qualitative examples and corresponding metrics are reported in Table 1.

Compared to the high quality of generated MNIST samples, the metrics on T4096 are generally low (e.g. 0.047 density), while the metrics on R64 report reasonably high scores (e.g. 0.491 density). For sound data, likewise, the metrics on T4096 do not faithfully represent the general fidelity and diversity of sound samples. The real and fake sound samples are provided at <http://bit.ly/38DIMAA> and <http://bit.ly/2HAM8NB>. The samples consist of human voice samples of digit words “zero” to “nine”. The fake samples indeed lack enough fidelity yet, but they do cover diverse number classes. Thus, the recall (0.029) and coverage (0.020) values under T4096 are severe underestimations of the actual diversity. Under the R64, the recall and coverage values are in the more sensible range: 0.572 and 0.653, respectively. When the target data domain significantly differ from the embedding training domain, R64 may be a more reasonable choice.

5. Conclusion and discussion

We have systematically studied the existing metrics for evaluating generative models with a particular focus on the fidelity and diversity aspects. While the recent work on the improved precision and recall (Kynkäänniemi et al., 2019) provide good estimates of such aspects, we discover certain failure cases where they are not practical yet: overestimating the manifolds, underestimating the scores when the real and fake distributions are identical, not being robust to outliers, and not detecting certain mode dropping. To remedy the issues, we have proposed the novel density and coverage metrics. Density and coverage have an additional conceptual advantage that they allow a systematic selection of involved hyperparameters. We suggest future researchers to use density and coverage for more stable and reliable diagnosis of their models. On top of this, we analyse the less-studied component of embedding. Prior metrics have mostly relied on ImageNet pre-trained embeddings. We argue through empirical studies that random embeddings are better choices when the target distribution is significantly different from the ImageNet statistics.

Acknowledgements

We thank the Clova AI Research team for the support, in particular Sanghyuk Chun and Jungwoo Ha for great discussions and paper reviews. Naver Smart Machine Learning (NSML) platform (Kim et al., 2018) has been used in the experiments. Kay Choi has advised the design of figures.

References

- Bengio, Y., Mesnil, G., Dauphin, Y., and Rifai, S. Better mixing via deep representations. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.
- Borji, A. Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding*, 2019.
- Brock, A., Donahue, J., and Simonyan, K. Large scale gan training for high fidelity natural image synthesis. *International Conference on Learning Representations*, 2018.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
- Donahue, C., McAuley, J., and Puckette, M. Adversarial audio synthesis. In *International Conference on Learning Representations*, 2019.
- Dowson, D. and Landau, B. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.
- Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pp. 6626–6637, 2017.
- Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- Kim, H., Kim, M., Seo, D., Kim, J., Park, H., Park, S., Jo, H., Kim, K., Yang, Y., Kim, Y., et al. Nsml: Meet the mlaas platform with a real-world case study. *arXiv preprint arXiv:1810.09957*, 2018.
- Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, pp. 3929–3938, 2019.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *International Conference on Learning Representations*, 2016.

Sajjadi, M. S., Bachem, O., Lucic, M., Bousquet, O., and Gelly, S. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, pp. 5228–5237, 2018.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. In *Advances in neural information processing systems*, pp. 2234–2242, 2016.

Simon, L., Webster, R., and Rabin, J. Revisiting precision recall definition for generative modeling. In *International Conference on Machine Learning*, pp. 5799–5808, 2019.

Theis, L., van den Oord, A., and Bethge, M. A note on the evaluation of generative models. In *International Conference on Learning Representations*, pp. 1–10, 2016.

Torralba, A. and Efros, A. A. Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9446–9454, 2018.

Vaserstein, L. N. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Perekhodchikov Informatsii*, 1969.

Yu, F., Zhang, Y., Song, S., Seff, A., and Xiao, J. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.