Topology Distance: A Topology Based Approach for Evaluating Generative Adversarial Networks

Danijela Horak,[1] **Simiao Yu,** [1] **Gholamreza Salimi-Khorshidi** [1,2]
[1] AIG
[2] University of Oxford
danijela.horak@aig.com, simiao.yu@aig.com, reza.khorshidi@aig.com

Abstract

Automatic evaluation of the goodness of Generative Adversarial Networks (GANs) has been a challenge for the field of machine learning. In this work, we propose a distance complementary to existing measures: Topology Distance (TD), the main idea behind which is to compare the geometric and topological features of the latent manifold of real data with those of generated data. More specifically, we build Vietoris-Rips complex on image features, and define TD based on the differences in persistent-homology groups of the two manifolds. We compare TD with the most commonly-used and relevant measures in the field, including Inception Score (IS), Fréchet Inception Distance (FID), Kernel Inception Distance (KID) and Geometry Score (GS), in a range of experiments on various datasets. We demonstrate the unique advantage and superiority of our proposed approach over the aforementioned metrics. A combination of our empirical results and the theoretical argument we propose in favour of TD, strongly supports the claim that TD is a powerful candidate metric that researchers can employ when aiming to automatically evaluate the goodness of GANs' learning.

Introduction

Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) are a class of deep generative models that have achieved unprecedented performance in generating high-quality and diverse images (Brock, Donahue, and Simonyan 2019). They have also been successfully applied to a variety of image-generation tasks, e.g. super resolution (Ledig et al. 2017), image-to-image translation (Zhu et al. 2017), and text-to-image synthesis (Reed et al. 2016), to name a few. The GAN framework consists of a generator G and a discriminator D, where G generates images $\mathrm{X_g}$ that are expected to resemble real images $\mathrm{X_r}$, while D discriminates between $\mathrm{X_g}$ and $\mathrm{X_r}$. G and D are trained by playing a two-player minimax game in a competing manner. Such novel adversarial training process is a key factor in GANs' success: It implicitly defines a learnable objective that is flexibly adaptive to various complicated image-generation tasks, in which it would be difficult or impossible to explicitly define such an objective.

One of the biggest challenges in the field of generative models – including for GANs – is the automatic evaluation of the goodness of such models (e.g., whether or not the data generated by such models are similar to the data they were trained on). Unlike supervised learning, where the goodness of the models can be assessed by comparing their predictions with the actual labels, or in some other deep-learning models where the goodness of the model can be assessed using the likelihood of the validation data under the distribution that the real data comes from, in most state of the art generative models we do not know this distribution explicitly or can not rely on labels for such evaluations.

Given that the data (or their corresponding features) in such situations can be assumed to lie on a manifold embedded in a high dimensional space (Goodfellow, Bengio, and Courville 2016), tools from topology and geometry come as a natural choice when studying differences between two data set. We propose topology distance (TD) for the evaluation of GANs; it compares the the topological structures of two manifolds, and calculates a distance between them to evaluate their (dis)similarities. We compare TD with widely-used and relevant metrics, and demonstrate that it is more robust to noise compared to competing distance measures on GAN's, and it is better suited to distinguish among various shapes that the data might come in. TD is able to evaluate GANs with new insights different from other existing measurements. It can therefore be used either as an alternative to, or in conjunction with other metrics.

Related Work

There have been multiple metrics proposed to evaluate the performance of GANs. In this paper we focus on the most commonly-used and relevant approaches (as follows); for a detailed account see (Borji 2018; Sajjadi et al. 2018).

Inception Score (IS) The main idea behind IS (Salimans et al. 2016) is that generated images of high quality are expected to meet two requirements: They should contain easily classifiable objects (i.e. the conditional label distribution $p(\mathrm{y}|\mathbf{x})$ with low entropy) and should be diverse (i.e. the marginal distribution $p(\mathrm{y})$ with high entropy). IS measures the average KL divergence between these two distributions:

$$\mathrm{IS} = \exp(\mathbb{E}_{\mathbf{x}\sim p_\mathrm{g}}[\mathrm{KL}(p(\mathrm{y}|\mathbf{x}) \mid\mid p(\mathrm{y}))]), \qquad (1)$$

where p_g is the generative distribution. IS relies on a pre-trained Inception model (Szegedy et al. 2016) for the classification of the generated images. Therefore, a key limitation of IS is that it is unable to evaluate the image types that are distinct from those that the Inception model was trained on.

Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) Proposed by (Heusel et al. 2017), FID relies on a pretrained Inception model, which maps each image to a vector representation (or, features). Given two groups of data in this vector space (one from the real and the other from the generated images), FID measures their similarities, assuming that the features are distributed as multivariate Gaussian; the distance will be the Fréchet distance (also known as Wasserstein-2 distance) between the two Gaussians:

$$\mathrm{FID}(p_\mathrm{r}, p_\mathrm{g}) = \left\|\mu_\mathrm{r} - \mu_\mathrm{g}\right\|_2^2 + \mathrm{Tr}(\Sigma_\mathrm{r} + \Sigma_\mathrm{g} - 2(\Sigma_\mathrm{r}\Sigma_\mathrm{g})^{\frac{1}{2}}) \quad (2)$$

where p_r and p_g denote the feature distributions of real and generated data, (μ_r, Σ_r) and (μ_g, Σ_g) denote the means and covariances of the corresponding feature distributions, respectively. It has been shown that FID is more robust to noise (of certain types) than IS (Heusel et al. 2017), but its assumption of features following a multivariate Gaussian distribution might be an oversimplification.

A similar metric to FID is KID (Mikołaj Bińkowski 2018), which computes the squared maximum mean discrepancy (MMD) between the features (learned also from a pretrained Inception model) of real and generated images:

$$\begin{aligned} \mathrm{KID}(p_\mathrm{r}, p_\mathrm{g}) =& \mathbb{E}_{\mathbf{x}_\mathrm{r}, \mathbf{x}'_\mathrm{r} \sim p_\mathrm{r}}[k(\mathbf{x}_\mathrm{r}, \mathbf{x}'_\mathrm{r})] \\ &+ \mathbb{E}_{\mathbf{x}_\mathrm{g}, \mathbf{x}'_\mathrm{g} \sim p_\mathrm{g}}[k(\mathbf{x}_\mathrm{g}, \mathbf{x}'_\mathrm{g})] \\ &- 2\mathbb{E}_{\mathbf{x}_\mathrm{r} \sim p_\mathrm{r}, \mathbf{x}_\mathrm{g} \sim p_\mathrm{g}}[k(\mathbf{x}_\mathrm{r}, \mathbf{x}_\mathrm{g})] \end{aligned} \quad (3)$$

where k denotes a polynomial kernel function $k(\mathbf{x}, \mathbf{x}') = (\frac{1}{\mathrm{d}}\mathbf{x}^\mathrm{T}\mathbf{x}' + 1)^3$ with feature dimension d. Compared with FID, KID does not have any parametric form assumption for feature distribution, and has a unbiased estimator.

Our proposed TD is closely related to FID and KID in that it also measures the distance between latent features of real and generated data. However, the key distinction of TD is that the target distance is computed by considering the geometric and topological properties of those latent features.

Geometry Score (GS) GS (Khrulkov and Oseledets 2018) Geometry score is defined as l_2 distance between means of the relative living-times (RLT) vectors associated with the two sets of images. RLT of a point cloud data (e.g., a group of images in the feature space) is an infinite vector $(\mathrm{u}_1, \mathrm{u}_2, \ldots)$ whose i-th entry is a measure of persistent intervals having 1-persistent homology group rank equal to i. That is, $\mathrm{u}_\mathrm{i} = \frac{1}{\mathrm{d}_{\mathrm{n}(\mathrm{n}-1)/2}} \sum_{\mathrm{j}=1}^{\mathrm{n}(\mathrm{n}-1)/2} \mathrm{I}_\mathrm{j}(\mathrm{d}_{\mathrm{j}+1} - \mathrm{d}_\mathrm{j})$, where I_j equals 1 if the rank of persistent homology group of dimension 1 in interval $[\mathrm{d}_\mathrm{j}, \mathrm{d}_{\mathrm{j}+1}]$ is i, and zero otherwise. Persistent homology parameters d_i, $\mathrm{i} \in [0 \ldots \mathrm{n}(\mathrm{n}-1)/2]$ are sorted distances in the observed point cloud data.

Geometry score exploits a similar idea to the TD, with the difference being in the underlying point cloud data used, dimensionality of the homology group and distance measure between the persistent diagrams. We claim that our method better aligns with the existing theory in the area of computational algebraic topology and has superior experimental results.

Main Idea

According to the manifold hypothesis (Goodfellow, Bengio, and Courville 2016), real world high dimensional data (and their features) lie on a low dimensional manifold embedded in a high dimensional space. The main idea of this paper is to compare the latent manifold of the real data with that of the generated data, based entirely on the topological properties of the data samples from these two manifolds. Let $\mathbb{M}_\mathrm{r}$ and $\mathbb{M}_\mathrm{g}$ be the latent manifolds of the real and generated data, respectively. We aim to compare these two manifolds using the finite samples of points V_r from $\mathbb{M}_\mathrm{r}$ and V_g from $\mathbb{M}_\mathrm{g}$.

Most mainstream methods compare samples V_r and V_g using the lower order moments (e.g. (Heusel et al. 2017)) – similar to the way we compare two functions using their Taylor expansion, for instance. However, this would only be valid if the underlying manifold is an Euclidean space (zero curvature), as all moments of the samples are calculated using Euclidean distance. For a Riemannian manifold with a nonzero curvature, this type of approach, at least in theory, would not work, and using geodesic instead of Euclidean distance would agree more with the hypothesis.

Here we propose the comparison of the two manifolds on the basis of their topology and/or geometry. The ideal way to compare two manifolds would be to infer if they are geometrically equivalent, i.e. isometric. This, unfortunately, is not attainable. However, we could compare two manifolds by the means of eigenvalues of the Laplace-Beltrami operator[1] on them.

The Laplace-Beltrami spectrum can be regarded as the set of squared frequencies that are associated to the modes of eigenvalues of an oscillating membrane defined on the manifold. The spectrum then, is an infinite sequence of eigenvalues, and satisfies some nice stability properties, whereby a small perturbation in the metric of the underlying Riemannian manifold results in a small perturbation of the spectrum (Donnelly 2010; Birman 1963). Furthermore, the Laplace-Beltrami spectrum is widely considered as a "fingerprint" of a manifold. In 1966, in the famous paper "Can one hear the shape of a drum?" (Kac 1966), M. Kac has asked a question whether the eigenvalues of Laplace Beltrami operator alone are sufficient to uniquely (up to an isometry) identify a manifold. The answer is unfortunately not, but the isospectral manifolds are rare and when they exist, they share multiple topological and geometric features.

Furthermore, it is possible to translate this methodology to a discrete setting, such that the spectrum calculated on the discrete set relates closely to the spectrum on the manifold itself.

Theorem 1 (Mantuano 2005) *Given a discretisation* $\mathrm{G} = \mathrm{G}(\mathrm{V}, \epsilon)$ *of a compact Riemannian manifold* $\mathbb{M}$ *which has non-negative sectional curvature* κ*, and non negative injectivity radius, and for which* $\mathit{Ricci}(\mathbb{M}, \mathrm{g}) \geq -(\mathrm{n}-1)\kappa\mathrm{g}$*, where* n *is dimension of a manifold,* g *is a Riemannian metric, then it is possible to associate the eigenvalues of Laplace operator on a graph* G*, with the ones of the Laplace Beltrami operator on* $\mathbb{M}(\mathrm{c}_1\lambda_\mathrm{k}(\mathrm{G}) \leq \lambda_\mathrm{k}(\mathbb{M}) \leq \mathrm{c}_2\lambda_\mathrm{k}(\mathrm{G}))$*, for*

[1] Only for Riemannian manifolds

all $k < |V|$.

The discretisation $\mathrm{G}(\mathrm{V},\epsilon)$ of a manifold $\mathbb{M}$, is a set of points in $\mathbb{M}$ whose distance is at least ϵ and the union of the balls centred in the points of V with radius ϵ which forms an open cover of $\mathbb{M}$, denoted by $\mathcal{U}$. A version of this theorem also holds for eigenvalues of higher dimensional version of Laplace Beltrami operator, called Laplace–de Rham operator which reflects high dimensional topological and geometric properties of a manifold (Mantuano 2008). This effectively means that for a sufficiently good sample V from $\mathbb{M}$, we can claim that calculating the eigenvalues on V would effectively be as calculating them on $\mathbb{M}$ (see (Dey, Ranjan, and Wang 2010) for more results).

In our case the manifold $\mathbb{M}$ is unknown, and all we know is a sample of points from it: V. In order to calculate the Laplace-Beltrami spectrum, we need to have a graph structure on V, which comes through Čech complex on its cover $\mathcal{U}$. To obtain the Čech complex, one needs radii of the balls in the cover, i.e., ϵ. This in itself poses a problem, because it is difficult to determine the right value of ϵ. There is little hope in recovering spectral properties of $\mathbb{M}$ from the point sample V, because we are unable to determine the right value of ϵ.

A similar theorem to Theorem 1 applies to homology type of a manifold and its sample.

Theorem 2 *Given a Riemannian manifold* $\mathbb{M}$, *and a sample of points from it,* V, *which is sufficiently dense, then a Vietoris–Rips complex of* V *at scale* ϵ *is homologically equivalent to* $\mathbb{M}$.

This theorem is a direct consequence of the famous nerve theorem (Alexandroff 1928), but can also be seen as a consequence of Theorem 1, due to a fact that the multiplicity of eigenvalue zero on discretised space is exactly the rank of a homology group of dimension zero on the same spacee.

In practice, as before, one does not know how to choose scale for ϵ, but unlike before, in this setting we have available a tool that can, and is specifically designed to, deal with the uncertainty of scale: persistent homology. We chose to utilise persistent homology to extract information about the geometry and topology of $\mathbb{M}$, because persistent homology, measured on the sample V, is a reliable shape quantifier of $\mathbb{M}$.

Preliminaries

Intuitively speaking, topological space is any space on which the notion on neighbourhood can be defined. Hence, all metric spaces (and consequently all examples considered in this work) are topological spaces; the opposite is not true (i.e., not all topological spaces can be endowed with a metric).

It is very difficult to directly assess whether two topological spaces are equivalent (homeomorphic); instead topologists use proxies to measure their similarity. One of these proxies are homology groups, denoted by $\mathrm{H}_\mathrm{k}, \mathrm{k} \in \mathbb{N}_0$, which loosely speaking encode the information on different types of loops (of different dimensions) that can be observed in the topological space. In this work we will only be concerned with a special class of topological spaces called simplicial complexes. Simplicial complex, commonly denoted by K, is a topological space consisting of vertices in a set V and a set of faces chosen from the partitive set of V, $\mathcal{P}(\mathrm{V})$, with the requirement that if $\mathrm{W} \in \mathrm{K}$, then all the subsets of W are also in K. One way to visualise the simplicial complex is to consider vertices as points in $\mathbb{R}^\mathrm{n}$ and m-dimensional faces as convex hulls of $\mathrm{m}+1$ vertices, i.e. edges, triangles, tetrahedra, etc.

Homology groups are algebraic constructions defined by

$$\mathrm{H}_\mathrm{k}(\mathrm{K},\mathbb{R}) = \mathrm{Z}_\mathrm{k}(\mathrm{K},\mathbb{R})/\mathrm{B}_\mathrm{k}(\mathrm{K},\mathbb{R}), \tag{4}$$

where k is a non-negative integer, $\mathrm{Z}_\mathrm{k}(\mathrm{K},\mathbb{R})$ is a vector space of k-dimensional cycles and $\mathrm{B}_\mathrm{k}(\mathrm{K},\mathbb{R})$ a k-dimensional boundaries, obtained as per-images and images of the boundary mapping on a chain complex, for more detailed account see (Hatcher 2009). Typically a rank of a homology group of dimension zero would be the number of connected components of K, rank of H_1 would be the number of one dimensional holes in K, rank of H_2 would be the number of cavities, and so on.

A persistent homology, is a homology of a topological space measured at ”different resolutions”. More precisely, we study a nested sequence of topological spaces (i.e., filtration $\mathcal{K} : \mathrm{K}_0 \subset \mathrm{K}_1 \subset \ldots \subset \mathrm{K}_\mathrm{N}$) and measure (calculate) homology at every step. As an example let’s observe the sequence of Vietoris-Rips complexes on a point set V: A Vietoris-Rips complex on a vertex set V and a diameter ϵ is a simplicial complex,in which $\mathrm{v}_0,\ldots,\mathrm{v}_k$ is a simplex iff $d(\mathrm{v}_\mathrm{i},\mathrm{v}_\mathrm{j}) < \epsilon$ for every $\mathrm{i},\mathrm{j} \leq \mathrm{k}$.

An example of a filtration would be $\mathcal{VR} : \mathrm{VR}(\mathrm{V},\epsilon_0) \subseteq \mathrm{VR}(\mathrm{V},\epsilon_1) \subseteq \ldots \subseteq \mathrm{VR}(\mathrm{V},\epsilon_\mathrm{n})$, where $0 = \epsilon_0 \leq \epsilon_1 \leq \ldots \leq \epsilon_\mathrm{n}$ (see Figure 1 (top)). In other words, persistent homology quantifies a change of topological invariants in VR with a change of parameter ϵ.

Formally,

$$\mathrm{H}_\mathrm{k}^{\mathrm{i},\mathrm{j}}(\mathcal{VR}) = \mathrm{Z}_\mathrm{k}^\mathrm{i}/\mathrm{B}_\mathrm{k}^\mathrm{j} \cap \mathrm{Z}_\mathrm{k}^\mathrm{i}, \tag{5}$$

where the j-th persistent homology group of dimension k of the i-th filtration complex $\mathrm{VR}(\mathrm{V},\epsilon_\mathrm{i})$ is denoted by $\mathrm{H}_\mathrm{k}^{\mathrm{i},\mathrm{j}}(\mathcal{VR})$. Intuitively, the persistent homology group records the “cycles” at the filtration step i, which have not become “boundaries” (i.e. which have not effectively disappeared) at filtration step j. For detailed account see (Edelsbrunner and Harer 2008).

The main insight when it comes to persistent homology is that the evolution of topological invariants over increase in parameter ϵ, can be encoded compactly in the form of a persistent diagram $\mathcal{PD}$ and a barcode.

$$\mathcal{PD}_\mathrm{k}(\mathcal{VR}) = \{(\mathrm{b}_\mathrm{i},\mathrm{d}_\mathrm{i}) \mid \mathrm{i} \in \mathbb{N}, \mathrm{b}_\mathrm{i},\mathrm{d}_\mathrm{i} \in \{\epsilon_0,\ldots,\epsilon_\mathrm{n}\}\}, \tag{6}$$

where b_i in the pair $((\mathrm{b}_\mathrm{i},\mathrm{d}_\mathrm{i}))$ records the appearance (or birth) of a k-dimensional homology group and d_i records its disappearance (also referred to as “death”). In the event that homology group persists, i.e. it does not disappear during the end of filtration, we set $\mathrm{d}_\mathrm{i} = \infty$. This set of points is represented in the upper triangle of the first quadrant of the $\mathbb{R}^2$ (see Figure 1 (bottom)). Another, representation is a barcode where each bar is mapped to a point $((\mathrm{b}_\mathrm{i},\mathrm{d}_\mathrm{i}))$ with the starting point b_i and ending point d_i.

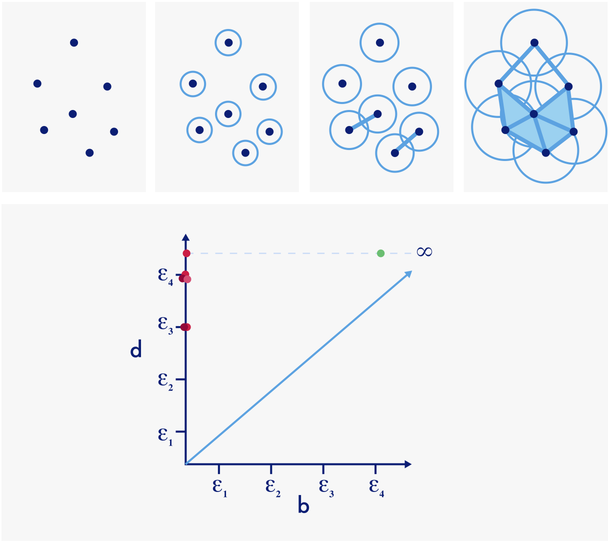

Figure 1: (top) The 4 step filtration of Vitoris Rips complex on the set of 7 points with increasing radius $0 = \varepsilon_1 \leq \ldots \leq \varepsilon_4$. (bottom) The persistent diagram corresponding to the filtration in the figure on top: b-axis denotes "birth" or appearance of persistent homology group, and d-axis denotes "death" or disappearance. Red points represent persistent homology groups of dimension 0, and the green ones of dimension 1.

There is a natural measure of distance defined on persistent diagrams, ∞-Wasserstein distance, also known in the community as the bottleneck distance, which has desirable stability properties with respect to small perturbations (Chazal et al. 2016), but is sensitive to outliers and mostly unsuitable for use in practice. On the other hand, p-Wasserstein distance

$$\mathrm{W_q}(\mathrm{L_p})(\mathcal{PD}_k^0, \mathcal{PD}^1) = \Big[\inf_{\eta:\mathcal{PD}_k^0 \to \mathcal{PD}_k^1} \sum_{\mathrm{u} \in \mathcal{PD}_k^0} \| \mathrm{u} - \eta(\mathrm{u}) \|_{\mathrm{q}} \Big]^{1/\mathrm{p}}, \tag{7}$$

where $1 \leq \mathrm{p}, \mathrm{q} \leq \infty$, and η ranges over all bijections between sets of persistent intervals in diagrams $\mathcal{PD}_0$ and $\mathcal{PD}_1$, shows more potential as presented in (Chazal et al. 2018), but is computationally demanding.

In practice, much of the applications of persistent homology have used neither of the two distances, but have relied on ad-hoc distances between persistent diagrams, which do not have a strong backing in theory (e.g. (Bendich et al. 2016; Khrulkov and Oseledets 2018). Endowed with any distance measure described above, the space of persistent diagrams is a metric space.

Method

The method we propose for evaluation of the performance of generative models rests on measuring the differences between the set of images generated by GANs and set of original images. We measure the distance on the point cloud data in feature space. Let $\mathrm{F_r}$ be the set of features of the real, and $\mathrm{F_g}$ of the generated images represented in: $\mathbb{R}^{\mathrm{m}}$.

Seen as the point cloud data in $\mathbb{R}^{\mathrm{m}}$, one can calculate the distances between the points in $\mathrm{F_r}$. It is worth noting here again, that even though we calculate all the distances using Euclidean metric, the algorithm will effectively use only "small" distances, and this is in agreement with potential non-zero curvature of the manifold(refer to Theorem 1 and Theorem 2 for full statement of this fact).

Assume that there are n data points in $\mathrm{F_r}$ and $\mathrm{F_g}$, and let $0 = \mathrm{d}_0^{\mathrm{r}} \leq \mathrm{d}_1^{\mathrm{r}} \leq \ldots \leq \mathrm{d}_{\mathrm{n(n-1)/2}}^{\mathrm{r}}$ and $0 = \mathrm{d}_0^{\mathrm{g}} \leq \mathrm{d}_1^{\mathrm{g}} \leq \ldots \leq \mathrm{d}_{\mathrm{n(n-1)/2}}^{\mathrm{g}}$ be an array of sorted distances among vectors in $\mathrm{F_r}$, $\mathrm{F_g}$, respectively. Then we observe the following filtration: $\mathcal{VR}(\mathrm{F_r}) : \mathrm{VR}(\mathrm{F_r}, \mathrm{d}_0^{\mathrm{r}}) \subseteq \ldots \subseteq \mathrm{VR}(\mathrm{F_r}, \mathrm{d}_{\mathrm{k}}^{\mathrm{r}})$ and $\mathcal{VR}(\mathrm{F_g}) : \mathrm{VR}(\mathrm{F_g}, \mathrm{d}_0^{\mathrm{g}}) \subseteq \ldots \subseteq \mathrm{VR}(\mathrm{F_g}, \mathrm{d}_{\mathrm{l}}^{\mathrm{g}})$, where the distance $\mathrm{d}_{\mathrm{k}}^{\mathrm{r}}$ is the minimal distance d for which corresponding Vietoris Rips complex becomes fully connected. Same is true for $\mathrm{d}_{\mathrm{l}}^{\mathrm{r}}$. The 0 dimensional persistent homology groups are calculated on the aforementioned filtrations. One consequence of studying only 0th dimensional persistent homology group, is that the rank of the persistent homology group at time 0 will be exactly n, and persistent diagram will consist of n pairs $(\mathrm{b_i}, \mathrm{d_i})$, where $\mathrm{b_i}$ denotes the point in filtration where the observed homology group has appeared for the first time (In our case $\mathrm{b_i} = 0$, for every i, due to the choice of filtration), and $\mathrm{d_i}$ denotes a point in filtration where the observed homology group(connected component) has merged with another one, or is equal to ∞ otherwise. This observation holds for both $\mathrm{F_r}$ and $\mathrm{F_g}$.

A commonly used distance between persistence diagrams in the field of TDA is bottleneck distance. However,in addition to being sensitive to outliers bottleneck distance effectively ignores noise, and while this might be beneficial in some cases, in our use case is the opposite: the noise is one of the main indicators of image quality and must not be ignored by the distance metric. Hence, we've used the inherent properties of our filtration method to define distance metric suitable to the use case. We assign a n-dimensional vector $l(\mathrm{F_r}) = (\mathrm{d_0} - \mathrm{b_0}, \ldots, \mathrm{d_n} - \mathrm{b_n})$, called the **longevity vector** to the persistent diagram which represents the sorted living times of each homology group for point set $\mathrm{F_r}$. Same for $\mathrm{F_g}$.

We define the Topology Distance (TD) between two persistent diagrams, and consequently between two corpuses of images to be $\mathrm{l_2}$ distance between their longevity vectors, i.e. $\mathrm{TD}(\mathrm{F_r}, \mathrm{F_g}) = \| l(\mathrm{F_r}) - l(\mathrm{F_g}) \|_2$, where $l(\mathrm{F_r})$ and $l(\mathrm{F_g})$ are the longevity vectors of persistent diagrams of filtrations of set of original and generated image features, respectively.

As some persistent pairs may contain ∞ we will assume that the difference between two infinite coordinates in 0, and the difference between ∞ and non-infinite coordinate in our algorithm is a some fixed value larger than the maximum finite longevity. The TD algorithm is summarised in Algorithm 1 and Algorithm 2.

Experiments

Datasets and Experimental Setup We compared our proposed TD (lower is better) with IS (higher is better), FID (lower is better), KID (lower is better) and GS (lower is bet-

Algorithm 1 This algorithm is to compute the longevity vector l for a set of images. l is of length n, which represents living times of all n homology classes throughout filtration (see Method section for more details).

```
**Require:** $f^*_\theta$: a pretrained feature extractor with parameters $\theta$.
**Require:** $RC(\mathrm{p})$: a function for computing Vietoris-Rips Filtration over the given input points $\mathrm{p}$.
**Require:** $PD(\mathrm{c})$: a function for computing persistent homology in dimension 0 of filtration $\mathrm{c}$.
**Input:** $\mathrm{X} \in \mathbb{R}^{\mathrm{N}\times\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$: a set of images of size $\mathrm{H} \times \mathrm{W}$ with $\mathrm{C}$ channels.
**Compute:** $\mathrm{F} \leftarrow f^*_\theta(\mathrm{X})$
**Compute:** $\mathrm{C} \leftarrow RC(\mathrm{F})$
**Compute:** $l \leftarrow PD(\mathrm{C})$
```

Algorithm 2 This algorithm is to compute Topology Distance (TD) between real and generated images.

```
**Input:** $\mathrm{X_r} \in \mathbb{R}^{\mathrm{N}\times\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$: a set of real images of size $\mathrm{H} \times \mathrm{W}$ with $\mathrm{C}$ channels.
**Input:** $\mathrm{X_g} \in \mathbb{R}^{\mathrm{N}\times\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$: a set of generated images of size $\mathrm{H} \times \mathrm{W}$ with $\mathrm{C}$ channels.
**Compute:** $l_\mathrm{r}$ with $\mathrm{X_r}$ using Algorithm 1.
**Compute:** $l_\mathrm{g}$ with $\mathrm{X_g}$ using Algorithm 1.
**Compute:** $TD(\mathrm{X_r}, \mathrm{X_g}) \leftarrow \| l_\mathrm{r} - l_\mathrm{g} \|_2$
```

ter) as introduced in Related Work section. In addition to some simulated data, which we will introduce in the next Section, our experiments were carried out on the following four datasets: Fashion-MNIST (Xiao, Rasul, and Vollgraf 2017), CIFAR10, corrupted CIFAR100 (CIFAR100-C) (Hendrycks and Dietterich 2019) and CelebA (Liu et al. 2015). Wherever features were required for computing the metric, we used a ResNet18 model (He et al. 2016) trained from scratch for Fashion-MNIST images, and the Inception model (Szegedy et al. 2016) pretrained on ImageNet (Deng et al. 2009) for all other datasets.

We implemented our algorithm using Python version of GUDHI[1] for topology-related computation and PyTorch (Paszke et al. 2019) for building and training neural network models.

Comparison with FID and KID The idea of basing the distance measure entirely on the first two moments (e.g., *a la* FID) can be an oversimplification of the underlying distributions at times, as describing certain distributions require the use of higher order statistics (e.g., third or fourth moments). Furthermore, if two distributions have identical moments of all orders, it is still possible for them to be different distributions (Romano and Siegel 1986). This leads to a conclusion that any distance metric based entirely on moments cannot successfully distinguish between all probability distributions.

In order to assess how such theoretical considerations will affect FID score's performance, we first compared TD and FID on a synthetic dataset. As shown in Figure 2 we aim to calculate the distance between a single Gaussian distribution and a mixture of two Gaussian distributions (the mixture has the same mean and variance as the single Gaussian). Given the identical first and second moments of the two point clouds in this case, as expected, FID cannot discriminate between the two, whereas the difference is very obvious when using TD. KID has similar limitations as FID, as demonstrated in Figure 2.

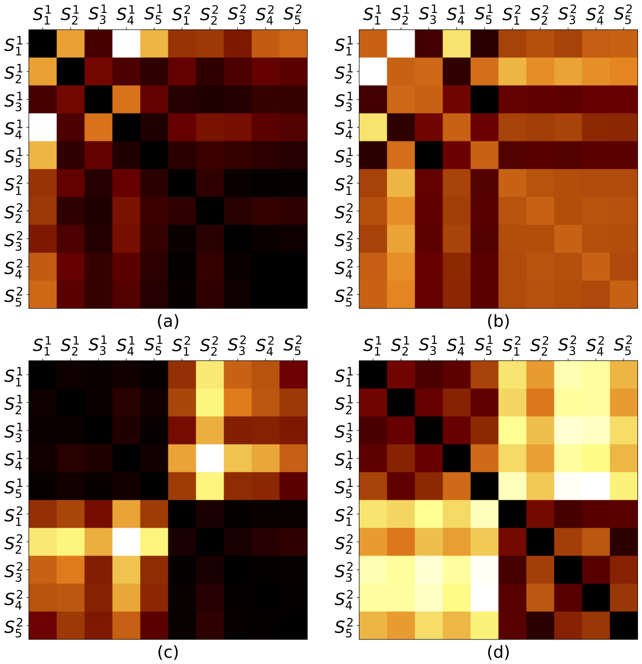

Figure 2: Heat maps of distance matrices between 5 sample sets (each of which has 600 samples randomly sampled from a single Gaussian distribution, denoted by s^1_i), and another 5 sample sets (each containing 600 samples from a mixture of two Gaussian distributions, denoted by s^2_i). All s^j_is have the same first and second moments. (a) FID. (b) KID. (c) GS. (d) TD.

Next, we compared TD with FID and KID on real images, randomly sampled from CelebA dataset; the goal was to compare the actual images with their manipulated counterparts. More specifically, we performed three types of manipulations designed by (Liu et al. 2018), which resulted in three new image datasets; we then computed the distance between each one of these manipulated image datasets and the original image dataset, using TD, FID and KID.

The three image manipulations include: 1) pixel noise (i.e., adding a random noise to each pixel, where the noise is uniformly sampled from the following interval: 1 ± 0.13 times the maximum intensity of the image), 2) patch mask (7 out of 64 evenly-divided regions of each image were masked by a random pixel from the image), and 3) patch exchange (2 out of 16 evenly-divided regions of each image were randomly exchanged with each other, performed twice). Some example images after manipulation are shown in Figure 3.

[1] http://gudhi.gforge.inria.fr/

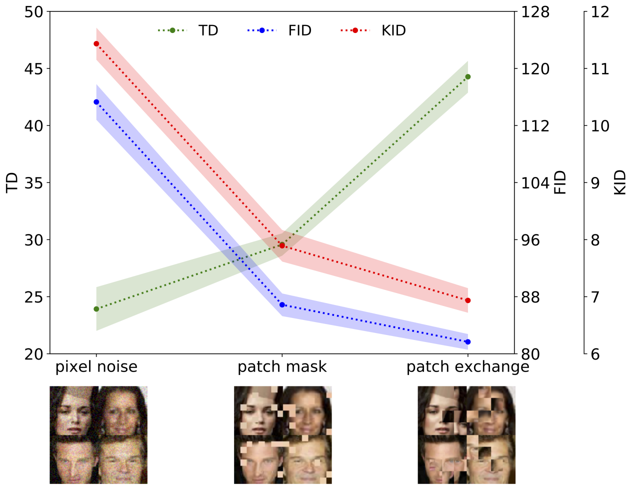

Figure 3: Comparison of FID, KID (×100) and TD on manipulated images (with pixel noise, patch mask and patch exchange). Results are averaged over 10 groups, each consisting of 500 real images (randomly sampled from the CelebA dataset) and the corresponding manipulated counterparts.

It is clear that the image quality increasingly worsens as we go from pixel noise, to patch mask and patch exchange; we expect to see this trend in the metrics.

However, FID and KID show a decreasing trend (indicating increasingly better quality) over pixel noise, patch mask and patch exchange, which is apparently opposite the human judgements. They fail to capture the worsening of image quality, as expected from qualitative assessments presented in (Liu et al. 2018); unlike TD, which captures the change in the quality of the manipulated images very well. Since all three metrics are based on the features extracted by the same Inception model, this experiment demonstrates that the superiority of TD over FID and KID is due to its effective assessment of the topological properties of the point clouds (rather than their lower-order statistics).

Comparison with GS So far, we have attempted to demonstrate the effectiveness of topology in assessing the (dis)similarities of two point clouds. On the other hand, both topology distance and geometry score exploit the idea of using topology to quantify dissimilarities between the latent manifolds of data. There are two major differences between TD and GS. The first one is in the core method and the way topology is used to construct the distance. The second one is that TD measures distances between point cloud data in the feature space, whereas geometry score is defined on raw pixels.

Figure 2c shows the heatmap of the distance matrix calculated between a single Gaussian distribution and a mixture of two Gaussian distributions using GS. It is clear that TD better discriminates between the samples from the two aforementioned distributions (see Figure 2d).

We then performed perturbation consistency comparison between TD and GS using the CIFAR100-C dataset, in which 16 different types of perturbations (grouped in four, namely noise, blur, weather and digital) are applied to the original CIFAR100 images; for each type of perturbation there are five levels of severity. 5,000 images were randomly sampled from the real dataset, and split into 10 groups (each with 500 images); for each group, scores are calculated comparing the perturbed images and the original. For every perturbation, as severity increases, the average score (across 10 groups) should increase monotonically with it.

As shown in Figure 4, TD is able to capture levels of perturbation severity much better and consistently than GS for many types of perturbations (e.g. Gaussian noise, frost, and elastic transform). This demonstrates that TD trend is more consistent with perturbation trend than GS, which further demonstrates the advantages of using features over pixels when computing topological properties.

Comparison with IS We then compared TD with IS on CelebA dataset where there are only face images and thus no distinct classes exist. We trained a GAN model (WGAN-GP (Gulrajani et al. 2017)) on the training set of CelebA; original images were cropped to be of size 64×64, and the model was then trained on them for 200 epochs with a batch size of 64. We recorded TD and IS along with the training process: every 4 epochs we fed the randomly sampled noise vector (remained fixed for different epochs) to the trained model so far; we then computed TD and IS on the generated and real images.

Figure 5 shows that TD has a great correspondence to the quality of generated images (i.e., decreasing trend with the improved quality of images). By contrast, IS fails to do so; at the early stages of training (before 20 epochs) it decreases as the quality of the generated images increases – in contrast to what is expected – and eventually loses its discrimination power at the remaining epochs. In summary, TD shows superiority over IS for evaluating the quality of images from datasets such as CelebA.

Pixels vs. Features Finally we performed an ablation study to compare the usage of pixels vs features when computing TD. We trained two WGAN-GP models, respectively, on Fashion-MNIST (trained for 100 epochs) and CIFAR10 (trained for 200 epochs) datasets. We then computed pixel-based and feature-based TD between images generated by WGAN-GP trained for different number of epochs and real images, randomly sampled from each dataset.

As can be seen clearly from Figure 6, for both datasets, feature-based TD is able to demonstrate better performance in terms of discrimination and consistency. This attributes to the better generalisation of learned features than raw pixels, which is one of the most significant advances of deep neural networks (Bengio, Courville, and Vincent 2013).

We also preformed a comprehensive comparison of image quality correlation among FID, KID, GS, IS and TD, and showed TD demonstrates better (vs. GS and IS) or comparable (vs. FID and KID) performance. The results can be found in the supplementary material.[1]

[1] https://arxiv.org/abs/2002.12054

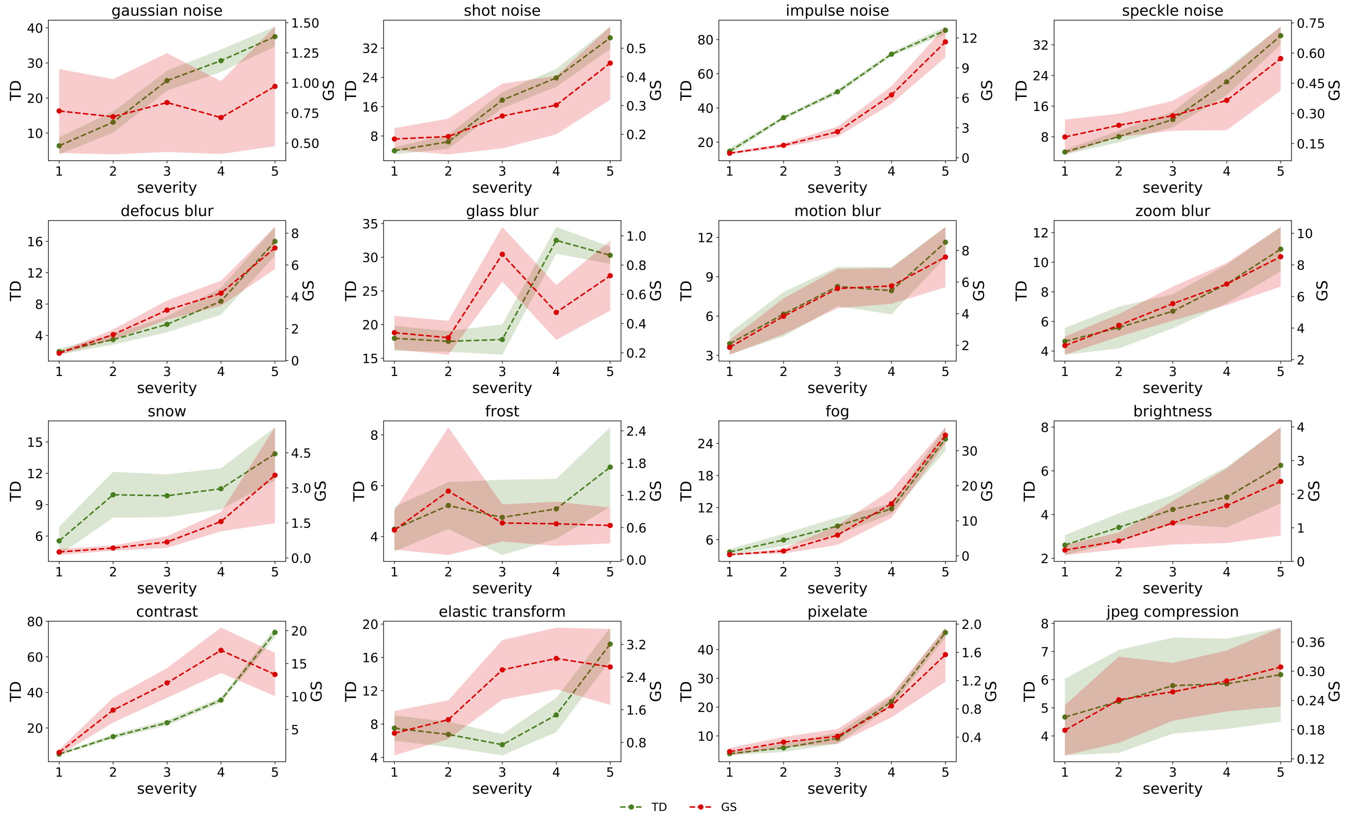

Figure 4: Comparison of perturbation consistency between TD and GS (×1000) on CIFAR100-C dataset. Results are averaged over 10 groups, each of which consists of 500 real images (randomly sampled from the CIFAR100-C dataset) and the corresponding images with perturbation.

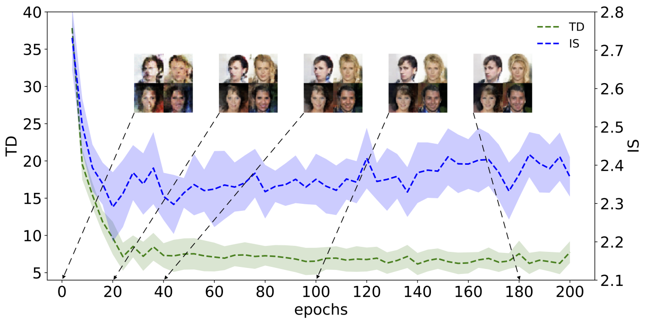

Figure 5: Comparison of TD and IS on generated images by WGAN-GP along with training process on the CelebA dataset. Results are averaged over 10 groups, each of which consists of 500 real images (randomly sampled from the original dataset) and generated images.

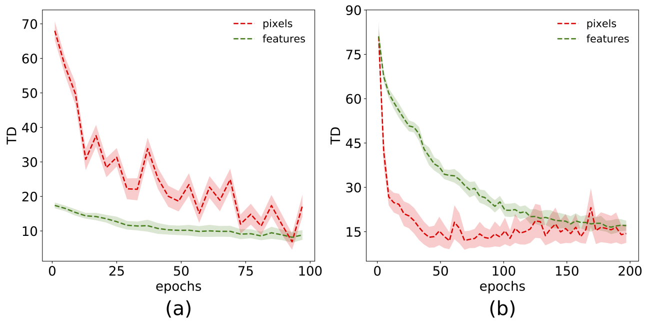

Figure 6: Comparison of TD based on pixels and features on different datasets. Results are averaged over 10 groups, each of which consists of 500 real images (randomly sampled from the original dataset) and generated images. (a) Fashion-MNIST. (b) CIFAR10.

Conclusion

In this work, we introduced Topology Distance (TD), a novel metric to evaluate GANs by considering the topological structures of latent manifold of real and generated images. In a range of experiments we have compared TD with IS, FID, KID, and GS, and have demonstrated its advantages and superiority over them in terms of consistency with human judgement, as well as other quantitative measures of change in image quality. TD is capable of providing new insights for the evaluation of GANs, and it thus can be used in conjunction with other metrics when evaluating GANs.

References

Alexandroff, P. 1928. Über den allgemeinen Dimensionsbegriff und seine Beziehungen zur elementaren geometrischen Anschauung. *Mathematische Annalen* 98: 617–635.

Bendich, P.; Marron, J. S.; Miller, E.; Pieloch, A.; and Skwerer, S. 2016. Persistent homology analysis of brain artery trees. *Ann. Appl. Stat.* 10(1): 198–218.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 35(8): 1798–1828.

Birman, M. S. 1963. On existence conditions for wave operators. *Izv. Akad. Nauk SSSR* 27(4): 883–906.

Borji, A. 2018. Pros and Cons of GAN Evaluation Measures. *arXiv preprint arXiv 1802.03446* .

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *ICLR*.

Chazal, F.; De Silva, V.; Glisse, M.; and Oudot, S. 2016. The Structure and Stability of Persistence Modules. In *Springer Briefs in Mathematics*.

Chazal, F.; Fasy, B.; Lecci, F.; Michel, B.; Rinaldo, A.; and Wasserman, L. 2018. Robust Topological Inference: Distance To a Measure and Kernel Distance. *J. Mach. Learn. Res.* 18(159): 1–40.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*.

Dey, T. K.; Ranjan, P.; and Wang, Y. 2010. Convergence, Stability, and Discrete Approximation of Laplace Spectra. In *SODA*.

Donnelly, H. 2010. Spectral Theory of Complete Riemannian Manifolds. *Pure Appl. MATH. Q.* 6(2): 439–456.

Edelsbrunner, H.; and Harer, J. 2008. Persistent homology — a survey. *Discrete Comput. Geom.* 453: 257.

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep Learning*. MIT Press.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In *NeurIPS*.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved Training of Wasserstein GANs. In *NeurIPS*.

Hatcher, A. 2009. *Algebraic Topology*. Cambridge University Press.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *ICLR*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *NeurIPS*.

Kac, M. 1966. Can One Hear the Shape of a Drum. *Amer. Math. Monthly* 73(4): 1–23.

Khrulkov, V.; and Oseledets, I. V. 2018. Geometry Score: a Method For Comparing Generative Adversarial Networks. In *ICML*.

Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Aitken, A. P.; Tejani, A.; Totz, J.; Wang, Z.; and Shi, W. 2017. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *CVPR*.

Liu, S.; Wei, Y.; Lu, J.; and Zhou, J. 2018. An Improved Evaluation Framework for Generative Adversarial Networks. *arXiv preprint arXiv 1803.07474* .

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *ICCV*.

Mantuano, T. 2005. Discretization of Compact Riemannian Manifolds Applied to the Spectrum of Laplacian. *Ann. Glob. Anal. Geom.* 27: 33–46.

Mantuano, T. 2008. Discretization of Riemannian manifolds applied to the Hodge Laplacian. *Am. J. Math.* 130(6): 1477–1508.

Mikołaj Bińkowski, Dougal J. Sutherland, M. A. A. G. 2018. Demystifying MMD GANs. In *ICLR*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS*.

Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; and Lee, H. 2016. Generative Adversarial Text to Image Synthesis. In *ICML*.

Romano, J. P.; and Siegel, A. 1986. *Counterexamples in Probability And Statistics*. Chapman and Hall/CRC.

Sajjadi, M. S. M.; Bachem, O.; Lucic, M.; Bousquet, O.; and Gelly, S. 2018. Assessing Generative Models via Precision and Recall. In *NeurIPS*.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved Techniques for Training GANs. In *NeurIPS*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the Inception Architecture for Computer Vision. In *CVPR*.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv 1708.07747* .

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *ICCV*.